ABSTRACT

Title of dissertation: Quantitative aspects of the stability of some dynamical systems

Cecilia I. González Tokman, Doctor of Philosophy, 2010

Dissertation directed by: Professor Brian R. Hunt
Department of Mathematics

This thesis is concerned with the study of quantitative aspects of the stability of some dynamical systems that exhibit hyperbolic features. Results include scaling laws for bubbling bifurcations, description of limit invariant measures for metastable systems and shadowing properties of a data assimilation algorithm in the context of hyperbolic systems.

Quantitative aspects of the stability of some dynamical systems

by

Cecilia I. González Tokman

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2010

Advisory Committee:

Professor Brian R. Hunt, Chair

Professor Dmitry Dolgopyat

Professor Michael Jakobson

Professor Vadim Kaloshin

Professor Edward Ott

A mis padres.

Acknowledgment

The research presented in this thesis has been supported by CONACyT, Mexico.

Table of Contents

List of Figures

Chapter 1

Introduction

This thesis is concerned with the study of quantitative aspects of the stability of certain dynamical systems that exhibit some type of strong hyperbolic features. The content of Chapters 2 and 3 is closely related to the the topic of open dynamical systems, which aims to understand the evolution of systems with holes in the phase space. This topic has received increasing attention in the last few years; see [DY06] for an exposition of such systems.

In Chapter 2, we establish rigorous scaling laws for the average bursting time for bubbling bifurcations of an invariant manifold, assuming the dynamics within the manifold to be hyperbolic. This type of global bifurcation appears in nearly synchronized systems, and is conjectured to be typical among those breaking the invariance of an asymptotically stable hyperbolic invariant manifold. We consider bubbling precipitated by generic bifurcations of a fixed point in both symmetric and non-symmetric systems with a codimension one invariant manifold, and discuss their extension to bifurcations of periodic points. We also discuss generalizations to invariant manifolds with higher codimension, and to systems with random noise. Most of this work was published in [GTH09], jointly with Brian R. Hunt.

In Chapter 3, we consider a piecewise smooth expanding map of the interval possessing two invariant subsets of positive Lebesgue measure and exactly two

ergodic absolutely continuous invariant probability measures (ACIMs). When this system is perturbed slightly to make the invariant sets merge, we describe how the unique ACIM of the perturbed map can be approximated by a convex combination of the two initial ergodic ACIMs. The result is generalized to the case of finitely many invariant components. This work, joint with Brian R. Hunt and Paul Wright, was recently accepted for publication, [GTHW].

The goal of Chapter 4 is to investigate a data assimilation procedure (DAP), an ensemble Kalman filter (EKF) studied in [HKS07], in the context of hyperbolic systems. We show that for every trajectory on an attractor, the predictions produced by the DAP remain close to the truth for all time provided the ensemble is properly initialized, making the DAP reliable. We deal with the case of one-dimensional unstable direction first, and later extend to higher dimensional unstable spaces. A feature of this approach is that no model linearizations are involved, making it efficient and potentially of interest for applications in high dimensional systems. Lyapunov exponents are also investigated.

Chapter 2

Scaling laws for bubbling bifurcations

2.1 Introduction

The goal of this article is to quantify how quickly an attracting invariant manifold with internally chaotic dynamics loses stability through a bubbling bifurcation in a certain class of systems. This type of bifurcation occurs when the invariant manifold ceases to be a asymptotically stable due to one of its embedded orbits becoming unstable in a direction transverse to the manifold [ABS96]. Under this circumstance, the invariant manifold can still attract a set of positive Lebesgue measure (and thus, support a physical measure). However, this attractor is extremely sensitive to small perturbations that make the manifold non-invariant. This scenario arises, for example, in physical systems with approximate (but not exact) symmetry, and can give rise to intermittent dynamics called *bubbling*, where a trajectory spends most of its time near the manifold but occasionally bursts away.

There are experimental results and formal calculations for particular models that predict scaling laws for the average time between bursts and the size of the perturbed attractor as a function of bifurcation parameters in generic bifurcation scenarios, see [VHO+96] and references therein. Our results make rigorous the theoretical predictions presented in [ZHO03], concerning scaling laws for the average interburst time in terms of parameters and positive Lyapunov exponents of the

bifurcating orbit. We prove the validity of similar scaling laws for more general dynamical systems displaying bubbling bifurcations. These results are applicable to generic systems, as well as to systems that have inherent symmetries.

The scaling laws we describe involve two parameters. One is a normal parameter as defined in [ABS96]. The normal parameter does not affect the invariant manifold nor the dynamics within it; but does affect the dynamics transverse to the manifold. The other parameter is a *symmetry-breaking* parameter that we call q, which when nonzero, perturbs trajectories from the manifold that is invariant for $q = 0$. An example with two such parameters is as follows.

$$(u_n, v_n) \mapsto (u_{n+1}, v_{n+1}) = (G(u_n, 0) + k(v_n - u_n), G(v_n, q) + k(u_n - v_n)). \quad (2.1)$$

Here, the invariant manifold for $q = 0$ is the synchronization manifold $u = v$. The coupling strength k is a normal parameter.

In this article, we consider a model of such systems in the form of skew-product as follows:

$$(x_n, y_n) \mapsto (x_{n+1}, y_{n+1}) = (T(x_n), F(x_n, y_n, p, q)), \quad (2.2)$$

where p and q are the parameters of the model. We consider x and y to be coordinates along and transverse to the invariant manifold, respectively, where we have made the simplifying assumption that the dynamics of x are independent of y and the parameters. We study the case of a uniformly hyperbolic base map T with an invariant SRB (or *physical*) measure μ.

For $q = 0$, we assume that the system has an invariant manifold $\{(x, y)|y = 0\}$, that the corresponding invariant measure $\mu \times \delta_0$ undergoes a bubbling bifurcation

at a fixed point $(x^*, 0)$ at $p = 0$ and that p is a *normal parameter* in the sense of [ABS96]. More generally, our results apply to systems that can be written in the form (2.2) by a choice of coordinates in a neighborhood of an invariant manifold. We discuss the scope of this model in §2.3.2. In more general cases, the orbit losing stability could be a periodic orbit. At various points in this paper we discuss how to extend our results to this case.

For $q = 0$, trajectories in the basin of $\mu \times \delta_0$ visit every neighborhood of $(x^*, 0)$. For definiteness, we assume that $(x^*, 0)$ is stable to perturbations of y for $p < 0$ and unstable for $p > 0$. The results of [ABS96] imply that generically, when $p > 0$ is sufficiently small the invariant measure $\mu \times \delta_0$ still has a basin of attraction with positive Lebesgue measure. For $p > 0$ and $q \neq 0$, trajectories that come close to $(x^*, 0)$ can *burst* away from $y = 0$. Depending on the y dynamics, trajectories that burst may come back close to $y = 0$ and repeat the bursting behavior, or they may remain away from $y = 0$. The former type of dynamics is called bubbling and the latter, transient dynamics. The existence of a physical or SRB measure for $p > 0$ and $q \neq 0$ and its dependence on parameters is a difficult question. Some results in this direction, in the context of partially hyperbolic diffeomorphisms, can be found in [Dol04] and references therein. Our results do not distinguish between bubbling and transient phenomena, and estimate the average time it takes for a trajectory initialized near $y = 0$ to burst for the first time. In the case of bubbling, we expect this average bursting time also to be representative of the average time between bursts. For the sake of exposition, we refer to the bifurcation that leads to bursts as a bubbling bifurcation, whether or not bubbling actually occurs.

The scaling of the average bursting time for small p and q depends on the type of bifurcation the fixed point $(x^*, 0)$ undergoes when $q = 0$ and p passes through 0. As in [ZHO03], we consider both the case of a generic transcritical bifurcation (Theorem 2.1) and that of a generic pitchfork bifurcation (Theorem 2.2). Our main results are stated in §2.2.2. We study two qualitatively different forms of bursting: *multiplicative* and *additive*. In the former case, bursts are driven by the dominant effect of the expansion parameter p. In the latter case, bursts occur due to the accumulation of perturbations to the system, quantified by q. (In [ZHO03], these were called drift-dominated and noise-dominated, respectively. We have adopted the new terminology to avoid possible confusion with other common uses of the former terms.) We also distinguish between hard and soft bifurcations. A hard bifurcation occurs when the maximum burst size changes suddenly as p increases, while in a soft bifurcation the maximum burst size increases gradually with p.

Besides providing a proof for the results predicted in [ZHO03], we extend the range of parameters over which the scaling law is valid, obtaining uniform bounds for the logarithm of the average bursting time, proportional to the sum of positive Lyapunov exponents of the bifurcating fixed point in the invariant manifold. Furthermore, we extend those results to more general dynamics: the scaling law is valid for skew-product systems with uniformly hyperbolic maps in the base variables (x) and for fiber (y) dynamics displaying a *generic* type of bifurcation explained in §2.2.1 (conditions (i)-(v)) and generalized in §2.3.2. These bifurcations include generic transcritical and pitchfork bifurcations of fixed points. Period-doubling bifurcations can also be treated with our tools, since the second power of a map

with a period-doubling bifurcation gives a map with a pitchfork bifurcation. Notice that a saddle-node bifurcation is not possible because the normal parameter assumption ensures that the fixed point persists on both sides of the bifurcation. Transcritical bifurcations can occur when the system is not symmetric with respect to reflection about the invariant manifold, while systems with reflectional symmetry will commonly have pitchfork bifurcations. As an example, the coupled system (2.1) is symmetric for $q = 0$, but if one of the coupling terms is eliminated it becomes asymmetric.

The structure of the paper is as follows. In Section 2.2, we first set up a model system in §2.2.1 and derive a Taylor approximation to F that we use in the subsequent sections. We state the main results in §2.2.2, and discuss three generalizations in §2.2.3; the case of a periodic bifurcating orbit in 2.2.3.1, the case of multiple transverse directions in 2.2.3.2, and a case of systems with random perturbations in 2.2.3.3. In Section 2.3 we analyze the dynamics and bifurcation of the invariant manifold. In §2.3.1, we prove some quantitative results about recurrence in hyperbolic systems that are relevant for our tasks. In §2.3.2, we discuss the mechanism of bubbling bifurcations and a generalization of the model presented in §2.2.1 to which our results apply. In Section 2.4 we prove the main results. In §2.4.1 we establish upper and lower bounds for the average bursting time in the linear regime (where the nonlinear terms in the Taylor approximation are small). In §2.4.2 we complete the proofs, extending those results to the nonlinear setting.

2.2 Statement of results

2.2.1 The model

Throughout this paper, we assume that we have a dynamical system with a compact, connected, hyperbolic invariant manifold X that undergoes a bubbling bifurcation. In our context, a bubbling bifurcation will be understood as the one that occurs when a parameter crosses a value at which the invariant manifold loses asymptotic stability. This loss of stability is due to one embedded orbit becoming unstable. In the terminology of [ABS96], at the bifurcation, the normal Lyapunov exponent to the invariant manifold X becomes 0 on one orbit but remains negative on other orbits.

To separate the dynamics on X from the transverse dynamics, we will work with skew-product systems: we assume that (near X) the dynamical system can be written in the form (2.2), where $T : X \circlearrowleft$ is a transitive C^2 Anosov diffeomorphism or a uniformly expanding map, and F is C^{1+1} in x and is C^3 as a function of y, p and q. We let μ be the SRB measure for T (see 2.3.1.1).

For $q = 0$ we impose the following conditions:

(i) $F(x, 0, p, 0) = 0$ for all x and p, so that $X = \{(x, y)|y = 0\}$ is an invariant manifold.

(ii) $x^* \in X$ is a fixed point. We let Λ be the sum of positive Lyapunov exponents of x^* associated to T.

(iii) X is asymptotically stable for $p < 0$, $\frac{\partial F}{\partial y}(x, 0, 0, 0) > 0$ for all x, $\frac{\partial F}{\partial y}(x^*, 0, 0, 0) =$

1 and $\frac{\partial^2 F}{\partial p \partial y}(x^*, 0, 0, 0) > 0$, so that $p = 0$ is a bifurcation value corresponding to the loss of asymptotic stability of X.

We remark that the assumption that $\frac{\partial F}{\partial y}(x, 0, 0, 0) > 0$ always holds if the map (2.2) is a diffeomorphism, because then $\frac{\partial F}{\partial y}(x, 0, 0, 0)$ must be nonzero for all x, and if it is negative we consider the second iterate of (2.2). The following assumption related to (iii) is generalized to the non-degeneracy condition (iii") in §2.3.2.

(iii') The global maximum of $\frac{\partial F}{\partial y}(\cdot, 0, 0, 0)$ is unique and occurs at x^*. This implies that the orbit $(x^*, 0, 0, 0)$ is the only orbit becoming unstable as p passes through 0.

(iv) The bifurcation of the fixed point x^* as p goes through 0 is either a generic transcritical bifurcation (in the asymmetric case) or a generic pitchfork bifurcation (in the symmetric case, where $F(x, y, p, 0) = -F(x, -y, p, 0)$).

We also assume the non-degeneracy condition:

(v) $\frac{\partial F}{\partial q}(x^*, 0, 0, 0) \neq 0$, so that varying q from 0 breaks the invariance of X near x^*.

With these requirements in mind, our model takes the following form:

$$\begin{cases} x_{n+1} &= T(x_n) \\ y_{n+1} &= F(x_n, y_n, p, q) \\ &= (f(x_n) + h(x_n)p)y_n + qg(x_n) + \mathcal{O}(qy + p^2 y + pq + q^2 + y^2), \end{cases} \tag{2.3}$$

where $f(x) = \frac{\partial F}{\partial y}(x, 0, 0, 0)$, $g(x) = \frac{\partial F}{\partial q}(x, 0, 0, 0)$, $h(x) = \frac{\partial^2 F}{\partial p \partial y}(x, 0, 0, 0)$. Notice that $\frac{\partial^k F}{\partial p^k}(x, 0, 0, 0) = 0$ for all $k \geq 1$ by condition (i) above.

For definiteness, we assume $q \geq 0$, and we think of q as the strength of the asymmetry in the system; we also refer to the term $qg(x)$ as the *kick*. By (iii) above, $f(x^*) = 1$ and $h(x^*) > 0$, and by the non-degeneracy condition (v), we have $g(x^*) \neq 0$. In fact, without loss of generality, we assume $g(x^*) = 1 = h(x^*)$. This amounts to possibly rescaling q and p, and possibly changing the sign of y.

For $p > 0$ and $q = 0$, the invariant manifold X is no longer asymptotically stable due to the fixed point x^* becoming unstable in a direction transverse to the manifold. However, since $0 < f(x) < 1$ for $x \neq x^*$, then most orbits close to X continue to be attracted to X. This is due to the fact that when p is small, the transverse dynamics is contracting outside a neighborhood of $x = x^*$.

Let $a(x) = \frac{1}{\rho!}\frac{\partial^\rho F}{\partial y^\rho}(x, 0, 0, 0)$, where $\rho \in \{2, 3\}$ corresponds to the most significant non-linearity of the dynamics of x^* for $q = 0$, that is, $\rho = 2$ for a transcritical bifurcation and $\rho = 3$ for a pitchfork bifurcation. Then $a(x^*) \neq 0$, and without loss of generality, we can rescale y to assume $a(x^*) = \pm 1$. The remaining higher order terms involve only higher powers of y, p and q.

If the system does not have inherent symmetry constraints, we have generically that $\rho = 2$, and the bifurcation that x^* goes through as p crosses 0 is a transcritical bifurcation. In this case, we can write:

$$F(x, y, p, q) = (f(x) + h(x)p)y + qg(x) + a(x)y^2 + \mathcal{O}(qy + p^2y + pq + q^2 + y^3), \quad (2.4)$$

with $a(x^*) \neq 0$.

On the other hand, if the system is symmetric with respect to changing the sign of y, or if x^* undergoes a period-doubling bifurcation and we consider the

second iterate of the map, the generic value is $\rho = 3$ and the corresponding generic bifurcation for x^* is a pitchfork bifurcation. In this case, we can write:

$$F(x, y, p, q) = (f(x) + h(x)p)y + qg(x) + a(x)y^3 + b(x)y^2 + \mathcal{O}(qy + p^2y + pq + q^2 + y^4), \tag{2.5}$$

with $a(x^*) \neq 0$ and $b(x^*) = 0$ (of course, $b(x) = 0$ for all x if $F(x, y, p, 0)$ is an odd function of y).

In both scenarios, the size of the bursts may be small and determined by the size of the perturbation parameters. We call this case a soft transition; it happens if $qa(x^*)g(x^*) < 0$ in the asymmetric case, and if $a(x^*) < 0$ in the symmetric case. If $qa(x^*)g(x^*) > 0$ in the asymmetric case or $a(x^*) > 0$ in the symmetric case, the size of bursts is not so limited; we call this case a hard transition.

2.2.2 Main results

In order to state the results, we introduce some notation. For a fixed threshold $Y > 0$ and $\{(x_n, y_n)\}_{n \in \mathbb{Z}_+}$ trajectory of (2.2), we define its bursting time as:

$$\tau(Y, x_0, y_0) = \min_{n \geq 0} \{|y_n| > Y\}.$$

Recall that μ is the SRB measure for $T : X \circlearrowleft$. For y_0 fixed, we define the average bursting time as:

$$\tau(Y, y_0) = \frac{1}{2y_0} \int_{X \times [-y_0, y_0]} \tau(Y, x, y) d\mu(x) dy.$$

Since perturbations from the invariant manifold $y = 0$ are proportional to q, we will generally consider y_0 to be of order q and set $\tau(Y) := \tau(Y, q)$. We will simply write τ when the threshold is clear from the context.

Remark 2.2.1. Our proofs also apply to the case where T is a nontransitive Anosov diffeomorphism or, more generally, an Axiom A diffeomorphism, with x^* belonging to a hyperbolic attractor $\mathcal{A}$. In this case, the basin of the SRB measure μ supported in $\mathcal{A}$ may no longer have full Lebesgue measure, and there may be other SRB measures for T, supported away from $\mathcal{A}$.

Our main result in the case of generic transcritical bifurcations ($\rho = 2$) is the following.

Theorem 2.1. Consider a family of skew product systems as in (2.3), with F as in (2.4) satisfying all conditions in § 2.2.1 above (2.3). Assume that $p, q > 0$. Then, there is a constant $\tilde{C} > 1$ and a threshold Y independent of p and q in the hard transition case ($qa(x^*)g(x^*) > 0$), and proportional to $\max(p, \sqrt{q})$ in the soft transition case ($qa(x^*)g(x^*) < 0$), such that the scaling of the bursting time satisfies:

- (Multiplicative case). For each $\epsilon > 0$, if $(p, \frac{q}{p^2})$ is sufficiently close to $(0, 0)$ and $q \geq p^2 e^{-p\tilde{C}^{\frac{1}{p}}}$, then
$$(1 - \epsilon)\Lambda < \frac{\log \tau(Y)}{\frac{1}{p} \log \frac{p^2}{q}} < (1 + \epsilon)\Lambda.$$
- (Additive case). There exists $C > 0$ independent of p, q and the map T on X such that for $(q, \frac{p^2}{q})$ sufficiently close to $(0, 0)$,
$$C^{-1}\Lambda \leq \frac{\log \tau(Y)}{\frac{1}{q^{\frac{1}{2}}}} \leq C\Lambda.$$

(Recall that Λ is the sum of positive Lyapunov exponents of the fixed point x^*.) This result is proved in §2.4.2.1.

In the case of pitchfork bifurcations, which are generic for symmetric systems ($\rho = 3$), the main result is:

Theorem 2.2. Consider a family of skew product systems as in (2.3), with F as in (2.5) satisfying all conditions in Section 2.2.1 above (2.3). Assume that $p, q > 0$. Then, there is a constant $\tilde{C} > 1$ and a threshold Y independent of p and q in the hard transition case ($a(x^*) > 0$), and proportional to $\max(\sqrt{p}, \sqrt[3]{q})$ in the soft transition case ($a(x^*) < 0$), such that the scaling of the bursting time satisfies:

- (Multiplicative case). For each $\epsilon > 0$, if $(p, \frac{q}{p^{\frac{3}{2}}})$ is sufficiently close to $(0, 0)$ and $q \geq p^{\frac{3}{2}} e^{-p\tilde{C}^{\frac{1}{p}}}$, then

$$(1 - \epsilon)\Lambda \leq \frac{\log \tau(Y)}{\frac{1}{p} \log \frac{p^{\frac{3}{2}}}{q}} \leq (1 + \epsilon)\Lambda,$$

- (Additive case). There exists $C > 0$ independent of p, q and the map T on X such that for $(q, \frac{p^{\frac{3}{2}}}{q})$ sufficiently close to $(0, 0)$,

$$C^{-1}\Lambda \leq \frac{\log \tau(Y)}{\frac{1}{q^{\frac{2}{3}}}} \leq C\Lambda.$$

This result is proved in §2.4.2.2.

The results predicted in [ZHO03], with the additional hypothesis that q is not exponentially small compared to p, are consequences of Theorems 2.1 and 2.2. These results are:

Corollary 2.3. Consider the following model systems of bubbling bifurcations:

$$\begin{cases} x_{n+1} &= 2x_n \pmod 1 \\ y_{n+1} &= (f(x_n) + p)y_n + ay_n^\rho + q \quad \text{for } |y| < 1 \text{ and } \rho \in \{2, 3\}, \end{cases} \tag{2.6}$$

where $f(0) = 1$, $0 < f(x) < 1$ for $x \neq 0$, $a \neq 0$, and parameters $p, q > 0$ are sufficiently small. In the multiplicative regime ($p^{\frac{\rho}{\rho-1}} \gg q > p^{\frac{\rho}{\rho-1}} e^{-p\tilde{C}^{\frac{1}{p}}}$, for some $\tilde{C} > 1$), for a threshold Y chosen as in the theorems above, the average bursting time obeys the following scaling laws:

$$\lim_{(p,\frac{q}{p^2})\to(0,0)} \frac{\log \tau(Y)}{\frac{1}{p}\log\frac{p^2}{|a|q}} = \log 2, \quad \text{when the coupling is asymmetric } (\rho = 2) \text{ and}$$

$$\lim_{(p,\frac{q}{p^{\frac{3}{2}}})\to(0,0)} \frac{\log \tau(Y)}{\frac{1}{p}\log\frac{p^{3/2}}{|a|^{1/2}q}} = \log 2, \quad \text{when the coupling is symmetric } (\rho = 3).$$

We have included in the conclusion of Corollary 2.3 terms from [ZHO03] involving a; while these terms do not affect the limits, they may make the limits converge faster.

Remark 2.2.2. The function $f(x) = \cos(2\pi x)$ considered in [ZHO03] does not meet our hypotheses because for technical reasons we have assumed f to be positive. However, our proofs can be adapted to such an f.

2.2.3 Generalizations

Here, we discuss three generalizations of our results. The first one concerns the replacement of the bifurcating fixed point by a periodic orbit. The second one is about the case of multidimensional transverse direction, that is, when the invariant manifold has codimension greater that one. The last one is to the case of random additive noise.

2.2.3.1 Periodic bifurcating orbit.

In case the bifurcating orbit is periodic of period d instead of a fixed point, after imposing non-degeneracy conditions, we could set up a model for bubbling bifurcations similar to that of 2.2.1. In this situation, when x gets near the periodic orbit, we would study the d-th power of T. The main difference with the fixed point case is that instead of having just one fixed point to keep track of, we would have d of them, and this introduces some technical difficulties. Although we do not carry out in detail all the calculations needed for this generalization, we do discuss the differences with the fixed point situation and provide ideas of how to extend the theorems in this case; see Remarks 2.3.7, 2.3.9 and 2.4.8.

2.2.3.2 Multidimensional transverse direction.

In this section we discuss a generalization of our analysis to the case when the codimension of the bifurcating invariant manifold X is larger than 1. As in hypotheses (i), (ii) and (iii) in §2.2.1, we assume that the attracting chaotic invariant manifold disappears when a direction transverse to X becomes unstable, that the orbit that first becomes unstable is a fixed point x^*, and that the bifurcation occurs at the parameter value $p = 0$.

Let $\vec{y}$ represent the multidimensional directions complementary to x. Our model system (2.3) then becomes:

$$\begin{cases} x_{n+1} & = T(x_n) \\ \vec{y}_{n+1} & = \vec{F}(x_n, \vec{y}_n, p, q). \end{cases}$$

If $\vec{F}$ is sufficiently smooth with respect to $\vec{y}$, p and q, we can bound the higher order terms as before. Following [ABS96], we impose the non-degeneracy condition that for $p = q = 0$, the fixed point x^* has a unique neutral direction transverse to X with eigenvalue 1, and that all other eigenvalues of $\frac{\partial \vec{F}}{\partial \vec{y}}(x^*, 0, 0, 0)$ have magnitude less than 1. We choose a norm defined by an inner product for $\vec{y}$, such that the corresponding norm of $\frac{\partial \vec{F}}{\partial \vec{y}}(x^*, 0, 0, 0)$ is equal to 1. Corresponding to (iii'), we assume that there are functions f and h on X, with f having a unique maximum of 1 at $x = x^*$, such that $\|\frac{\partial \vec{F}}{\partial \vec{y}}(x, 0, p, 0)\| = f(x) + h(x)p + \mathcal{O}(p^2)$. Then, one can show that the largest eigenvalue of $\frac{\partial \vec{F}}{\partial \vec{y}}(x^*, 0, p, 0)$ is $1 + h(x^*)p + \mathcal{O}(p^2)$. We call $\vec{v}(p)$ the corresponding eigenvector for the adjoint to $\frac{\partial \vec{F}}{\partial \vec{y}}(x^*, 0, p, 0)$, and let $\vec{g}(x) = \frac{\partial \vec{F}}{\partial q}(x, 0, 0, 0)$. Then, we can bound the growth of the norm of $\vec{y}$ as in the one-dimensional case,

$$|\vec{y}_{n+1}| \leq (f(x_n) + h(x_n)p)|\vec{y}_n| + q\|\vec{g}(x_n)\| + \mathcal{O}(q|\vec{y}_n| + p^2|\vec{y}_n| + pq + q^2 + |\vec{y}_n|^2).$$

Thus, the analysis of §2.4.1.2 remains applicable. We can bound the growth of the norm of $\vec{y}$ from below in a similar way, with an additional error term of order $|x_n - x^*||\vec{y}_n|$.

$$\begin{aligned} \vec{y}_{n+1} \cdot \vec{v}(p) \geq \quad & (f(x_n) + h(x_n)p)\vec{y}_n \cdot \vec{v}(p) + q\vec{g}(x_n) \cdot \vec{v}(p) \\ & + \mathcal{O}(q|\vec{y}_n| + p^2|\vec{y}_n| + pq + q^2 + |\vec{y}_n|^2 + |x_n - x^*||\vec{y}_n|). \end{aligned}$$

(If $\vec{g}(x^*) \cdot \vec{v}(0) < 0$, we change the sign of $\vec{v}$.) In order for the analysis in §2.4.1.1 to be applicable, the conditions that need to be satisfied are that $\vec{g}(x^*) \cdot \vec{v}(0) \neq 0$,

corresponding to (v), and, additionally, non-degeneracy conditions analogous to (iv) making x^* undergo a transcritical or pitchfork bifurcation. The main remaining complication to adapting the arguments in § 2.4 is that there is no analogue of Lemma 2.4.1 in this case, because the direction of $\vec{y}$ can rotate while x is away from x^*.

2.2.3.3 Random perturbations.

The results are generalized to some random dynamical systems, where the deterministic mismatch $g(x_n)$ is replaced by a stationary sequence of independent random variables g_n. This is presented in §2.5. We could also treat the case of combined random noise and deterministic mismatch with our methods.

2.3 Invariant manifold: dynamics and bifurcation

2.3.1 Dynamics on the invariant manifold

In this section, we present results we need for the base dynamics given by T, a transitive C^2 Anosov diffeomorphism. We assume T has a fixed point x^*, and derive quantitative dynamical properties that are used in our estimates in §2.4, following references [Bow75], [Che02] and [Aba04].

We note that all results of this section also apply to expanding maps. In particular, the model system with base dynamics given by $T(x) = mx \pmod 1$ is rich enough to give a good understanding of most of these properties. For general T, the analysis we present is somewhat more involved.

2.3.1.1 Existence of Markov partitions and SRB measures.

Classical works of Sinai [Sin68] and Bowen [Bow70] show that uniformly hyperbolic dynamical systems have Markov partitions of arbitrarily small diameter. Such partitions allow one to study the dynamics in symbolic terms, since all invariant measures of hyperbolic systems are projections of invariant measures on symbolic systems that are semi-conjugate to T.

Moreover, given any point $x \in X$, Pesin and Weiss show [PW97, Thm. 3] that, after possibly passing to a power of T, a Markov partition $\mathcal{R}$ can be chosen in such a way that x is in the interior of a Markov rectangle. We will choose such a *special* Markov partition with the hyperbolic fixed point x^* in the interior of a rectangle we call R_0.

Because of our hypotheses on T, there is always an invariant measure that is physically relevant: the SRB or physical measure [Sin68, Bow70]. We will call it μ. This measure is the one of interest for us, since its basin contains a full Lebesgue measure set of trajectories. Another relevant property of μ is exploited in §2.3.1.3, namely, that the μ measures of cylinders around a point (see definition below) are asymptotically determined by the sum of positive Lyapunov exponents.

For a fixed Markov partition $\mathcal{R} = \{R_0, \ldots, R_{D-1}\}$ of (X, T), we denote by $\omega_i(x)$ the index of the partition set to which $T^i(x)$ belongs, provided that $T^i(x)$ belongs to only one partition set. Note that this is undefined on the set for which $T^i(x)$ belongs to the boundary of a partition set, which has μ measure zero [Che02, Prop. 3.1]. We denote by $\Omega_{T,\mathcal{R}}$ the set of sequences $(\omega_i)_{i\in\mathbb{Z}} \subset \Sigma_D$ allowed by the

dynamics of T. Cylinder sets are nonempty sets $S \subset X$ of the form $S = \{x \in X | \omega_i(x) = b_i, k \leq i \leq k + l\}$, up to a set of μ measure zero, for some $k \in \mathbb{Z}$, $l \geq 0$ and $b_i \in \{0, 1, \ldots, D-1\}$. Such a cylinder set S has length $l+1$ and is based at k. We write $\mathcal{C}(k, k+l)$ to denote the collection of all cylinders of length $l+1$ and base k. We say that two cylinders $S_i \in \mathcal{C}(k_i, k_i')$, $i = 1, 2$, are determined by non-overlapping words if either $k_1' < k_2$ or $k_2' < k_1$.

Below, we use $\mathbb{P}_\mu$ to denote the probability of an event with respect to μ. For example,

$$\mathbb{P}_\mu(\omega_k = b) = \mu(\{x \in X | \omega_k(x) = b\}).$$

We also use $\mathbb{E}_\mu$ for the expectation with respect to μ.

2.3.1.2 Expected hitting time.

For a μ measurable set S with $\mu(S) > 0$, let $\tau_S(x)$ be the first time the orbit of x visits (or hits) S, that is, $\tau_S(x) = \min\{k \geq 0 | T^k(x) \in S\}$. By ergodicity, the *hitting time* $\tau_S(x)$ is finite for μ almost every $x \in X$ and defines a μ measurable function on X. The following lemmas relate the expected hitting time with $\mu(S)$. The first one, which follows from [Aba04, §5], gives an upper bound and holds for cylinder sets. The second one gives a lower bound and is valid for all measurable sets S of sufficiently small measure.

Lemma 2.3.1. There exists a constant $\tilde{U} = \tilde{U}(T, \mathcal{R}) \geq 1$ such that for every cylinder set S,

$$\mathbb{E}_\mu(\tau_S) := \int_X \tau_S(x) d\mu(x) \leq \frac{\tilde{U}}{\mu(S)}.$$

Lemma 2.3.2. If $\mu(S) < \frac{1}{4}$ then

$$\mathbb{E}_\mu(\tau_S) := \int_X \tau_S(x) d\mu(x) \geq \frac{1}{4\mu(S)}.$$

Proof. Let $S_k := \{x \in X | \tau_S(x) = k\}$. Hence $S_k \subseteq T^{-k}(S)$ and therefore $\mu(S_k) \leq \mu(S)$, which gives the lower bound

$$\begin{aligned}
\mathbb{E}_\mu(\tau_S) = \sum_{k=0}^{\infty} k\mu(S_k) = \sum_{k=1}^{\infty}\sum_{j=1}^{k} \mu(S_k) = \sum_{j=1}^{\infty}\sum_{k=j}^{\infty} \mu(S_k) = \sum_{j=1}^{\infty}\left(1 - \sum_{k=0}^{j-1} \mu(S_k)\right) \\
\geq \sum_{j=1}^{\lfloor \frac{1}{\mu(S)} \rfloor} (1 - j\mu(S)) = \lfloor \frac{1}{\mu(S)} \rfloor - \frac{\lfloor \frac{1}{\mu(S)} \rfloor (\lfloor \frac{1}{\mu(S)} \rfloor + 1)}{2} \mu(S) \\
= \lfloor \frac{1}{\mu(S)} \rfloor \left(1 - \frac{\mu(S)(\lfloor \frac{1}{\mu(S)} \rfloor + 1)}{2}\right) \geq \left(\frac{1}{\mu(S)} - 1\right)\frac{1-\mu(S)}{2} \geq \frac{1}{4\mu(S)},
\end{aligned}$$

where the last inequality follows from the fact that $\mu(S) < \frac{1}{4}$. □

2.3.1.3 Consecutive number of iterates near a fixed point.

It is necessary for our purposes to understand the distribution of the number of consecutive iterates a trajectory spends in a neighborhood of the fixed point x^*. Following traditional notation, we let

$$B_x(n, \epsilon) := \{z | \text{ dist}(T^i x, T^i z) < \epsilon \ \forall i = 0, 1, \ldots, n\}.$$

A trajectory x stays within ϵ of x^* for n iterates if $x \in B_{x^*}(n, \epsilon)$. Let $\Xi := \{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\dim X}\}$ be the Lyapunov spectrum of T at x^*. Let

$$\Lambda := \sum_{i=1}^{\dim X} (\lambda_i)_+, \text{ where } (\lambda)_+ := \max(\lambda, 0), \quad \text{and } \chi := e^\Lambda.$$

A lower bound on the number of iterates close to x^* is given by:

Lemma 2.3.3. There exist constants $C = C(T)$ and $\varphi = \varphi(\Xi)$ such that for all $\delta > 0$ sufficiently small,

$$\mu(B_{x^*}(n, \delta)) \geq C\delta^{\varphi}\chi^{-n}.$$

Proof. It is shown in [Bow75, § 4.4] that for $\epsilon > 0$ sufficiently small there is some constant $\tilde{C} = \tilde{C}(T, \epsilon)$ such that

$$\mu(B_{x^*}(n, \epsilon)) \geq \tilde{C}\chi^{-n}.$$

Also, by Corollary 6.4.17 in [KH95], if $\epsilon > 0$ is sufficiently small, there is a constant $A = A(T)$ such that if $x \in B_{x^*}(n, \epsilon)$ then for $0 \leq i \leq n$, $\operatorname{dist}(T^i x, x^*) \leq A\epsilon e^{-\tilde{\lambda}\min(i, n-i)}$, where $\tilde{\lambda} := \min_{i=1,\ldots,\dim X}(|\lambda_i|)$, the distance of Ξ to 0. Fix $\epsilon > 0$ and let $\varphi = 2\log\chi/\tilde{\lambda}$ and $k = \lceil \frac{\log\delta - \log A\epsilon}{\tilde{\lambda}} \rceil$. Then $A\epsilon e^{-\tilde{\lambda}k} \leq \delta$ and if $k \geq 0$ we have

$$\mu(B_{x^*}(n, \delta)) \geq \mu(B_{x^*}(n + 2k, \epsilon)) \geq \tilde{C}\chi^{-(n+2k)} \geq C\delta^{\varphi}\chi^{-n}. \qquad \square$$

Obtaining upper bounds for the time spent close to x^* requires a better understanding of the dynamical properties of T. Let $\xi_0(x), \xi_1(x), \ldots$ be the number of consecutive iterates the trajectory of x spends in successive visits to a Markov rectangle R_0 containing x^* in its interior.

The following lemmas will be useful in §2.4.1.

Lemma 2.3.4. There is a constant $A = A(T, \mathcal{R}) > 0$ such that for every $k \in \mathbb{N}$ and $t > 0$ we have:

$$\mathbb{P}_{\mu}(\xi_k \geq t) \leq A\chi^{-t}.$$

Proof. This is a consequence of the so-called exponential cluster property for uniformly hyperbolic systems (see e.g. [Bow75, Che02]): there are constants $\tilde{C}$ and

$\theta < 1$ such that given any two cylinders $S \in \mathcal{C}(0, a)$ and $S' \in \mathcal{C}(0, b)$,

$$|\mathbb{P}_\mu(S \cap T^{-n}(S')) - \mathbb{P}_\mu(S)\mathbb{P}_\mu(S')| \leq \tilde{C}\mathbb{P}_\mu(S)\mathbb{P}_\mu(S')\theta^{n-a}.$$

(Notice that $T^{-n}(S') \in \mathcal{C}(n, n+b)$, so $n-a$ represents the gap between the symbols determined by membership in S and those determined by membership in $T^{-n}(S')$.) In particular, there is a constant C such that for any two cylinders S and S' determined by non-overlapping allowed words, we have that S and S' are *independent up to a multiplicative factor* C in the following sense:

$$\mathbb{P}_\mu(S \cap S') \leq C\mathbb{P}_\mu(S)\mathbb{P}_\mu(S') \quad \text{and hence} \quad \mathbb{P}_\mu(S|S') \leq C\mathbb{P}_\mu(S),$$

where by $\mathbb{P}_\mu(S|S')$ we mean the conditional probability $\frac{\mathbb{P}_\mu(S \cap S')}{\mathbb{P}_\mu(S')}$.

To prove the lemma, we fix k and consider the following countable partition $\mathcal{Z}$ (modulo sets of μ measure 0) of $\Omega_{T,\mathcal{R}}$ as follows. Each element of $\mathcal{Z}$ consists of the cylinder set of sequences ω that share all symbols up to τ_k, where τ_k is the start of the k-th sequence of 0's. For example, for $k = 1$, sequences of the form $\mathbf{1}01\ldots$ and $\mathbf{1}001\ldots$ would belong to the same element of $\mathcal{Z}$, but sequences of the form $\mathbf{11}0\ldots$ would be in a different element of the partition. This is a partition modulo sets of μ measure 0 since with probability 1, $\tau_k < \infty$.

By the exponential cluster property, we know that for any $Z \in \mathcal{Z}$,

$$\mathbb{P}_\mu(\xi_k \geq t|Z) \leq C\mathbb{P}_\mu(0_t),$$

where 0_t is the sequence consisting of t zeros. Therefore,

$$\mathbb{P}_\mu(\xi_k \geq t) = \sum_{Z \in \mathcal{Z}} \mathbb{P}_\mu(\xi_k \geq t|Z)\mathbb{P}_\mu(Z) \leq C\mathbb{P}_\mu(0_t) \sum_{Z \in \mathcal{Z}} \mathbb{P}_\mu(Z) = C\mathbb{P}_\mu(0_t).$$

The lemma follows from the fact that for rapidly mixing systems, there exist positive constants $\lambda, \tilde{A}$ such that:

$$\mathbb{P}_\mu(0_t) < \tilde{A}e^{-\lambda t}.$$

Furthermore, since μ is an SRB measure, we can take $\lambda = \log \chi = \Lambda$ [Bow75]. $\square$

Given a constant $c < 1$, let $\eta_k := \sum_{j=0}^{k} c^{k-j}\xi_j$.

Lemma 2.3.5. For $c < 1$ fixed, there are constants B and $0 < \theta < 1$, such that whenever $B \leq t_* \leq t$,

$$\mu_k := \mathbb{P}_\mu(\eta_k \geq t,\ \eta_j \geq t_* \text{ for } j = 1, \ldots, k-1) \leq A\theta^k \chi^{-t}.$$

Proof. First, we show that for any values of t, t_j, the events $\{\xi_k = t\}$ and $\{\eta_j \geq t_j \text{ for } 0 \leq j < k\}$ are independent up to a multiplicative factor C given by the exponential cluster property. Let us make use of the partition $\mathcal{Z}$ from the proof of Lemma 2.3.4 and let $\mathcal{Z}_0 = \{Z \in \mathcal{Z} : \eta_j \geq t_j \text{ for } 0 \leq j < k\}$. We remark that $\mathcal{Z}_0$ is well defined, since for $0 \leq j < k$, η_j is constant μ-almost everywhere in each $Z \in \mathcal{Z}$. Thus, we have:

$$\begin{aligned}
\mathbb{P}_\mu(\xi_k = t, \eta_j \geq t_j \text{ for } 0 \leq j < k) &= \sum_{Z \in \mathcal{Z}} \mathbb{P}_\mu(\xi_k = t, \eta_j \geq t_j \text{ for } 0 \leq j < k | Z)\mathbb{P}_\mu(Z) \\
&= \sum_{Z \in \mathcal{Z}_0} \mathbb{P}_\mu(\xi_k = t|Z)\mathbb{P}_\mu(Z) \leq C\mathbb{P}_\mu(0_t)\mathbb{P}_\mu(\eta_j \geq t_j \text{ for } 0 \leq j < k) \\
&\leq CA\chi^{-t}\mathbb{P}_\mu(\eta_j \geq t_j \text{ for } 0 \leq j < k).
\end{aligned}$$

Now, using Lemma 2.3.4 as the base step, valid for all t_*, t, we will prove our result by induction. Assume we know that for some k, t_* and t we have that

$\mu_k \leq A\theta^k \chi^{-t}$. Then, for $k+1$ we have:

$$\begin{aligned}
\mu_{k+1} \quad &\leq \textstyle\sum_{s=0}^{t-ct_*} \mathbb{P}_\mu(\xi_{k+1} = s, \eta_k \geq \frac{t-s}{c}, \eta_j \geq t_* \text{ for } j = 0, \ldots, k) \\
&\quad + \mathbb{P}_\mu(\xi_{k+1} \geq t - ct_*, \eta_j \geq t_* \text{ for } j = 0, \ldots, k) \\
&\leq \textstyle\sum_{s=0}^{t-ct_*} C(A\chi^{-s})(A\theta^k \chi^{-(\frac{t-s}{c})}) + C(A\chi^{-(t-ct_*)})(A\theta^k \chi^{-t*}) \\
&\leq A\theta^k AC\left(\frac{\chi^{(\frac{1}{c}-1)}}{\chi^{(\frac{1}{c}-1)}-1} + 1\right)\chi^{-t}\chi^{-(1-c)t_*}.
\end{aligned}$$

This establishes the inductive step and the result provided

$$\theta = AC\left(\frac{\chi^{(\frac{1}{c}-1)}}{\chi^{(\frac{1}{c}-1)}-1} + 1\right)\chi^{-(1-c)B} \text{ and } B \text{ is large enough that } \theta < 1.$$

□

Let $N_k = \xi_0 + \xi_1 + \cdots + \xi_k$.

Lemma 2.3.6. For β sufficiently large and $0 < t \leq \chi^{\beta/2}$, there is a constant $0 < \tilde{\theta} < 1$ such that

$$\tilde{\mu}_k(t) := \mathbb{P}_\mu(N_k \geq t + k\beta, N_j \geq j\beta \text{ for } j = 0, \ldots, k-1) \leq A\tilde{\theta}^k \chi^{-t}.$$

Proof. By an argument analogous to the proof of Lemma 2.3.5, the events $\{\xi_k \geq t\}$ and $\{N_j \geq j\beta \text{ for } 0 \leq j < k\}$ are independent up to a multiplicative factor C. Using Lemma 2.3.4 as the base step, valid for all t, we will proceed by induction. The base step follows from Lemma 2.3.4. Assume we know that for some k and t $\tilde{\mu}_k(t) \leq A\tilde{\theta}^k \chi^{-t}$. For $k+1$ we have:

$$\begin{aligned}\tilde{\mu}_{k+1}(t) &\leq \sum_{s=1}^{\lfloor t+\beta \rfloor} C\mathbb{P}_{\mu}(\xi_{k+1}=s)\tilde{\mu}_k(t+\beta-s) + C\mathbb{P}_{\mu}(\xi_{k+1}>\lfloor t+\beta \rfloor)\tilde{\mu}_k(0) \\ &\leq A\tilde{\theta}^k CA\chi^{-(t+\beta)}\big(t+\beta+1\big) \leq A\tilde{\theta}^{k+1}\chi^{-t},\end{aligned}$$

provided $\tilde{\theta} := AC\big(t+\beta+1\big)\chi^{-\beta}$. For any choice of β, this last inequality gives explicit restrictions on the allowed size of t in order for $\tilde{\theta} < 1$. In particular, for sufficiently large β, the argument is valid for any $t \leq \chi^{\beta/2}$. □

Remark 2.3.7. The analysis of this section dealt with the dynamics close to a fixed point x^* of T. Results of this section can be adapted to study the dynamics close to a periodic orbit of period d after taking the d-th power of T. This extension is not completely trivial, since in this case we would have d fixed points $x_1^*, \ldots, x_d^*$ to simultaneously keep track of. However, straightforward extensions of the arguments in §2.3.1.3, yield similar bounds for the analogue of ξ_k, η_k, N_k in this setting. In this case, Λ would be replaced by the sum of positive Lyapunov exponents of the periodic orbit and $\chi = e^{\Lambda}$ would change accordingly.

2.3.2 Bifurcation of the invariant manifold

In this section, we discuss the genericity of conditions (iii) and (iii') of §2.2.1 about the bifurcation of the invariant manifold when $q = 0$. While condition (iii') is non-generic, we will weaken it to a condition (iii") that we characterize as a non-degeneracy assumption.

The assumption in condition (iii) that X is asymptotically stable for $p < 0$ implies that the Lyapunov exponent, equal to the average of $\log \frac{\partial F}{\partial y}$, is nonpositive for all invariant measures of T when $p < 0$. Recall that $f(x) = \frac{\partial F}{\partial y}(x, 0, 0, 0) > 0$, and observe that f is Lipschitz by our smoothness hypothesis for F. Then by continuity, the average of $\log f(x)$ is nonpositive for all invariant measures of T, and further (by condition (iii) again) the average is zero for the delta measure at x^*. Thus, among all invariant measures of T, the average of $\log f(x)$ is maximized at the bifurcating orbit. It has been conjectured [YH99] and numerically supported [HO96] that generically, maximizing (optimal) invariant measures occur at measures with periodic support. In this respect, we expect in general the loss of stability of X to occur at a periodic orbit, and for simplicity we consider the case of a fixed point x^*. Furthermore, it is a topologically generic property of Lipschitz and smooth functions [Jen06] to have a unique maximizing invariant measure. Condition (iii') makes the stronger assumption that pointwise the maximum of $\log f(x)$ occurs at x^*. We can easily weaken this assumption by requiring that it be true for some change of coordinates. Specifically, we consider

$$\tilde{y} = \eta(x)y, \quad \text{with } \eta(x) > 0 \text{ for all } x \in X \text{ and } \eta(x^*) = 1. \tag{2.7}$$

Under such coordinate change, the evolution equations for system (2.3) and parameters $p = q = 0$ become:

$$\begin{cases} x_{n+1} &= T(x_n) \\ \tilde{y}_{n+1} &= \tilde{F}(x_n, \tilde{y}_n, 0, 0), \end{cases} \tag{2.8}$$

with $\tilde{F}(x, \tilde{y}, 0, 0) = \eta(T(x))F(x, 0, 0, 0) + \tilde{y}\frac{\eta(T(x))}{\eta(x)}f(x) + \mathcal{O}(\tilde{y}^2)$. The corresponding

coefficient of the linear term in the Taylor expansion of $\tilde{F}$ with respect to $\tilde{y}$ becomes:

$$\tilde{f}(x) = \frac{\eta(T(x))}{\eta(x)} f(x).$$

Thus, condition (iii') can be replaced by

(iii") There exists a change of coordinates of the form (2.7) for which $\tilde{f}$ has a unique global maximum at x^*.

The following lemma suggests that it is plausible to expect such a change of coordinates.

Lemma 2.3.8. Let $f : X \to \mathbb{R}$ be a positive Lipschitz function. Suppose that among all T invariant measures, the average of $\log f$ is maximized at (the measure supported on) a fixed point x^*. Then, there exists a change of coordinates of the form (2.7) for which the global maximum of $\tilde{f}$ occurs at x^*.

Proof. Let $\phi(x) = \log f(x)$. This is well defined and Lipschitz, since f is positive and Lipschitz. Existence of a change of coordinates $\tilde{y} = \eta(x)y$ changing f into $\tilde{f}$ is equivalent to having a solution to the following co-homological equation:

$$\tilde{\phi}(x) = \phi(x) + \psi(T(x)) - \psi(x), \tag{2.9}$$

where $\tilde{\phi}(x) = \log \tilde{f}(x)$ and $\psi(x) = \log \eta(x)$.

When T is uniformly hyperbolic, the normal form theorem [Jen06, 4.7] ensures the existence of a Lipschitz solution ψ to (2.9) with the following property.

$$\phi(x^*) \geq \phi(x) + \psi(T(x)) - \psi(x) =: \tilde{\phi}(x).$$

Therefore, the change of coordinates from f to $\tilde{f}$ given by $\tilde{f}(x) = e^{\psi(x)} f(x)$ has a global maximum at x^*. □

With this result in mind, condition (iii") is similar to the assumption that the average of $\log f(x)$ over invariant measures of T has a unique maximum at x^*.

Remark 2.3.9. Lemma 2.3.8 extends to the case when the average of $\log f$ over the space of T invariant measures is maximized at a periodic orbit $x_1^*, \ldots, x_d^*$; after a coordinate change, the global maximum of f occurs at all d points of the orbit. In this case, our non-degeneracy assumption is that f is maximized only at these d points.

2.4 Proof of main results

All results in this section refer to dynamical systems of the form (2.3), satisfying assumptions (i)-(v) in §2.2.1 as well as either (iii') in §2.2.1 or (iii") in §2.3.2.

2.4.1 Average bursting time in the linear regime

The goal of this section is to derive a scaling law for the logarithm of the average bursting time τ, valid for burst amplitudes small enough that we can use a linear approximation to the y dynamics. We consider the effect of nonlinear terms in the following section. We set a threshold y value Y, and investigate the average time it takes for an initial condition starting close to X to burst (or *escape*) to the threshold.

When p is small, Lebesgue almost all orbits of T will spend most of their time in the region in which $f(x) + h(x)p < 1$, so that the y dynamics are contracting near $y = 0$. However, since x^* is in the support of μ, the x trajectory of Lebesgue

almost every orbit will visit arbitrarily small neighborhoods of x^* and thus remain close to x^* for arbitrarily long period of time, eventually resulting in a burst.

A quantitative understanding of this statement allows us to find a cylinder set $S \subset X$ such that whenever the x trajectory enters it, the trajectory is guaranteed to reach the threshold Y. From this, we obtain an upper bound for the average bursting time in terms of $\mu(S)$, since once in S, the time it takes to burst is relatively negligible.

The lower bound needs further work, since in order to establish it, an understanding of all possible escape routes to the threshold Y is needed. In this part, we will identify a set $S' \subset X$ (not necessarily a cylinder but a union of cylinders) such that the x coordinate of any trajectory that escapes must visit S' before escaping. The definition of the set S' depends on the fact that trajectories may escape not only through one long sequence of expansive iterates, but instead could follow a sequence of alternating expanding and contracting periods. We note that our results will show that the former is asymptotically the most likely escape route, provided q is bounded below as in the multiplicative cases of Theorems 2.1 and 2.2. The set S' also depends on an intermediate y threshold that is presented in §2.4.1.2.

In order to establish upper and lower bounds on the average bursting time, we restrict ourselves to finding lower and upper bounds on the measure of trajectories that initiate a burst, $\mu(S)$ and $\mu(S')$. This is enough for our purposes, in view of Lemmas 2.3.1 and 2.3.2.

We introduce two parameters for the threshold size: $\alpha = \frac{Y}{q}$ quantifies the number of iterates to reach the threshold for $x = x^*$ and $p = 0$, ignoring higher

order terms. The non-linearity parameter $s = \frac{Y^\rho}{\max(pY,q)}$ measures the size of the dominant non-linear term Y^ρ, relative to the largest term in the linearization of $y_{n+1} - y_n$ at $x_n = x^*$. Note that p, q and s determine Y and hence α.

Next, we bound the higher order terms in (2.3) by $\sigma p|y| + \zeta q$, where σ and ζ can be made arbitrarily close to 0 by making p, q and s small. In particular, we assume $\zeta, \sigma < 1$.

Throughout this section, we write $\tau = \tau(Y) = \tau(\alpha q)$, and recall that $\Lambda = \sum_{i=1}^{\dim X}(\lambda_i)_+ = \log \chi$ is the sum of positive Lyapunov exponents of x^* for T. We also recall that $q\Delta$ is upper bound on the *kick* $q(g(x) + \zeta)$. Let $0 < c < 1$ be an upper bound on $f(x) + h(x)p$ for $x \notin R_0$ and define

$$\tilde{l}(z) = \frac{1+z}{z\log(1+z)} \text{ and } K(p,\alpha) := \frac{1}{p\alpha\tilde{l}(p\alpha)} e^{\frac{p}{2}\chi^{\frac{\log\frac{1}{c}}{4p(1+\tilde{l}(p\alpha))}}}.$$

We say that parameters p, α satisfy condition $(\star)$ if α is sufficiently large (independent of p and q) and either $p\alpha < \frac{1}{2}\log\frac{1}{c}$, or $p\alpha \geq \frac{1}{2}\log\frac{1}{c}$ and $K(p,\alpha) > 1$. We remark that when $p\alpha$ is sufficiently large and p, q and s are sufficiently small, $K(p,\alpha) > 1$ if $k(p) := \frac{1}{p}e^{p\chi^{\frac{\log\frac{1}{c}}{5p}}} > \alpha$. The main result of this section is the following.

Theorem 2.4. For sufficiently small parameters p, q and s for which the threshold size α satisfies condition $(\star)$, there is a constant $C > 1$ independent of p, q and the map T on X such that

$$C^{-1}\Lambda \leq \frac{\log\tau(q\alpha)}{\frac{1}{p}\log(1+p\alpha)} \leq C\Lambda.$$

Moreover, in the limit that $\alpha \to \infty$ and either $p\alpha \to 0$ or $p\alpha \to \infty$, C can be taken arbitrarily close to 1.

In order to prove the theorem, we first show that there is a manifold $X_{-\tilde{\kappa}q} := X \times \{-\tilde{\kappa}q\}$, with $\tilde{\kappa} = \mathcal{O}(1)$ that gets mapped above itself (in the y direction), therefore preventing all initial conditions starting above it to escape the strip $X \times [-\alpha q, \alpha q]$ from below. Then, we establish the upper and lower bounds in §2.4.1.1 and §2.4.1.2, respectively.

Lemma 2.4.1. There is a constant $\tilde{\kappa}$ independent of p and q such that the image of any initial condition (x_0, y_0) of (2.3) with $y_0 \geq -\tilde{\kappa}q$ satisfies $y_1 \geq -\tilde{\kappa}q$ for all sufficiently small p and q. Moreover, in this case, there exists $\tilde{x} > 0$ independent of p and q such that every trajectory for which $y_0 > -\tilde{\kappa}q$ that remains in the set $|x - x^*| < \tilde{x}$ for a sufficiently long number of iterates n_0, independent of p and q, reaches a positive y value, that is $y_{n_0} > 0$.

Proof. Consider $\tilde{x}$ sufficiently small so that when $|x_0 - x^*| < \tilde{x}$ we have that $g(x_0) > \frac{1}{2}$, $h(x_0) < \frac{3}{2}$, and such that there is $0 < r < 1$ depending only on f such that for $|x_0 - x^*| \geq \tilde{x}$, $f(x_0) \leq 1 - 2r$. When p and q are sufficiently small, $1 + 2p \geq f(x_0) + (h(x_0) + \sigma)p$ for $|x_0 - x^*| < \tilde{x}$, and $0 \leq f(x_0) + (h(x_0) - \sigma)p \leq 1 - r$ for $|x_0 - x^*| \geq \tilde{x}$. Let $\tilde{\kappa} > \frac{1 - \min_{x \in X}\{g(x)\}}{r}$. For $|x_0 - x^*| < \tilde{x}$ and $y_0 \geq -\tilde{\kappa}q$, $y_1 \geq -q\tilde{\kappa}(1 + 2p) + (\frac{1}{2} - \zeta)q$. Hence, if p and q are sufficiently small, $\zeta < \frac{1}{4}$ and $y_1 \geq -\tilde{\kappa}q$. For $|x_0 - x^*| \geq \tilde{x}$ and $y_0 \geq -\tilde{\kappa}q$, we have $y_1 \geq -\tilde{\kappa}q(1 - r) + q(\min_{x \in X}\{g(x)\} - \zeta)$. If p and q are sufficiently small, $\zeta < 1$ and by the choice of $\tilde{\kappa}$, $y_1 \geq -\tilde{\kappa}q$.

For the second statement, we know that if $|x_0 - x^*| < \tilde{x}$ and $y_0 \geq -\tilde{\kappa}q$, then $y_1 - y_0 \geq (\frac{1}{2} - \zeta - 2p\tilde{\kappa})q$. The result follows from the fact that we can apply the estimate repeatedly, as long as $|x_i - x^*| < \tilde{x}$. □

2.4.1.1 Upper bound for the bursting time.

In this section, we take an initial condition (x_0, y_0) starting above the manifold $X \times \{-\tilde{\kappa}q\}$ and find a neighborhood of $x = x^*$ so that whenever the trajectory remains in it for a sufficiently long number of iterates, it is guaranteed to escape.

First, we present a simple upper bound useful in the additive case. Another upper bound will be obtained in Proposition 2.4.3 by taking into account the expansiveness close to x^*.

Proposition 2.4.2. For any $\epsilon > 0$, if p, q and s are sufficiently small, and α is sufficiently large, we have:

$$\frac{\log \tau}{\alpha} < (1+\epsilon)\Lambda.$$

Proof. Assume $L-1$ is a Lipschitz constant for $|f|+|g|+|h|$. Let $0 < \tilde{\tilde{\delta}} < 1$, $\epsilon' > 0$ sufficiently small and $\tilde{\tilde{x}} = \min\{\frac{1-\tilde{\tilde{\delta}}}{L}, \frac{\tilde{\tilde{\delta}}\epsilon'}{L\alpha}\}$. Then, for p, q and s sufficiently small, if $|x - x^*| < \tilde{\tilde{x}}$ we have $g(x) - \zeta > \tilde{\tilde{\delta}} > 0$ and $f(x) + (h(x) - \sigma)p > 1 - L\tilde{\tilde{x}} \geq 1 - \frac{\tilde{\tilde{\delta}}\epsilon'}{\alpha}$. In this situation, for $|x_n - x^*| \leq \tilde{\tilde{x}}$ and $y_n \leq \alpha q$ we have:

$$y_{n+1} \geq \big(f(x_n) + h(x_n)p\big)y_n - \sigma p|y_n| + q\big(g(x_n) - \zeta\big) \geq y_n + (1-\epsilon')\tilde{\tilde{\delta}}q.$$

Therefore, a trajectory starting with a positive y value reaches the threshold if it stays in the region $|x - x^*| < \tilde{\tilde{x}}$ for at least $\frac{\alpha}{(1-\epsilon')\tilde{\tilde{\delta}}} < \frac{\alpha(1+2\epsilon')}{\tilde{\tilde{\delta}}} =: \tilde{\tilde{n}}$ consecutive iterates. Thus, in this setting, we can take $S = B_{x^*}(\tilde{\tilde{x}}, \tilde{\tilde{n}} + n_0)$ as a *surely escaping* set, where $n_0 = \mathcal{O}(1)$ is as in Lemma 2.4.1. By Lemmas 2.3.1 and 2.3.3, it follows that there is a constant $U = U(T)$ such that for sufficiently large α, an upper bound on the

logarithm of the average bursting time is

$$\log \tau \leq \Lambda \frac{(1+3\epsilon')\alpha}{\tilde{\tilde{\delta}}} + \log U - \log \tilde{\tilde{x}}.$$

Since $\epsilon' > 0$ may be arbitrarily small and $\tilde{\tilde{\delta}}$ can be made arbitrarily close to 1 for p, q and s sufficiently small, the statement follows. □

Proposition 2.4.3. For any $\epsilon > 0$, if p, q and s are sufficiently small and $p\alpha$ is sufficiently large, an upper bound on the logarithm of the average bursting time is given by:

$$\frac{\log \tau}{\frac{1}{p}\log(1+p\alpha)} < (1+\epsilon)\Lambda.$$

Proof. Using the Lipschitz assumptions on f, g and h, we can find $\tilde{x} > 0$ to be specified later, and $\tilde{\delta} > 0$ so that for $|x - x^*| < \tilde{x}$ we have $f(x) + (h(x) - \sigma)p > 1 + \tilde{\gamma}p$ for some $0 < \tilde{\gamma} < 1 - \sigma$ and $g(x) - \zeta > \tilde{\delta} > 0$. By the choices of $\tilde{\kappa}$ and $\tilde{x}$, every trajectory with initial condition $y_0 \geq -\tilde{\kappa}q$ that is in the region $|x - x^*| < \tilde{x}$ for n_0 iterates we have that $y_{n_0} > 0$. Hence, if the trajectory remains in the region $|x - x^*| < \tilde{x}$ for another iterate, we will have that

$$\begin{aligned} y_{n_0+1} &\geq (f(x_{n_0}) + h(x_{n_0})p)y_n + g(x_{n_0})q - \sigma|py_{n_0}| - \zeta q \\ &\geq (1+\tilde{\gamma}p)y_{n_0} + \tilde{\delta}q, \end{aligned}$$

and if $|x - x^*| < \tilde{x}$ for another n consecutive iterates,

$$y_{n_0+n} \geq (1+\tilde{\gamma}p)^n \frac{\tilde{\delta}q}{\tilde{\gamma}p} - \frac{\tilde{\delta}q}{\tilde{\gamma}p}.$$

Hence, all orbits that remain in the region $|x - x^*| < \tilde{x}$ for time

$$\tilde{n} := n_0 + \frac{\log \frac{\tilde{\gamma}p\alpha + \tilde{\delta}}{\tilde{\delta}}}{\log(1+\tilde{\gamma}p)}$$

will reach the threshold αq within $\tilde{n}$ steps.

Thus, in this setting, we can take $S = B_{x^*}(\tilde{x}, \tilde{n})$ as a *surely escaping* set. By Lemma 2.3.3, we know that there are constants $C = C(T)$ and $\varphi = \varphi(T)$ such that:

$$\mu(S) \geq C\tilde{x}^{\varphi}\chi^{-\tilde{n}}.$$

Thus, by Lemma 2.3.1 we have that there is a constant $\tilde{U} = \tilde{U}(T)$ such that an upper bound on the average bursting time $\tau = \tau(Y)$ is:

$$\tau \leq \tilde{U}\tilde{x}^{-\varphi}\chi^{\tilde{n}} + \tilde{n}. \tag{2.10}$$

Therefore, there is a constant $U = U(T)$ such that for any $\epsilon > 0$, if p, q and s are sufficiently small and α sufficiently large, we have:

$$\log \tau \leq \left(1 + \frac{\epsilon}{2}\right)\Lambda\left(n_0 + \frac{\log \frac{\tilde{\gamma}p\alpha + \tilde{\delta}}{\tilde{\delta}}}{\log(1 + \tilde{\gamma}p)}\right) + \log U(1 - \log \tilde{x}).$$

Furthermore, we can make $\tilde{\gamma}$ and $\tilde{\delta}$ arbitrarily close to 1 by making $\frac{\tilde{x}}{p}, p, q$ and s sufficiently small and $p\alpha$ sufficiently large. Choosing $\tilde{x} = \frac{p}{\log(1+p\alpha)}$, the proposition follows. □

2.4.1.2 Lower bound for the bursting time.

To get a lower bound for the bursting time, we need to consider different *escape routes*. For a given y_0, in order for a trajectory starting at height less than $\tilde{y}_0 q$ to escape, it needs to get total expansion by a factor of $\frac{\alpha}{\tilde{y}_0}$. This expansion can be achieved in one long sequence of expansive iterates, which corresponds to the case presented in the previous subsection, or in several expansive sequences.

An important characteristic of our model is that the linearized contraction rate between any two expansive sequences is bounded above by some factor $0 < c < 1$ independent of p and α, and that it takes a long time to recover from it. These consideration will allow us to show that the measure of initial conditions that initiate an escape is comparable to the measure of initial conditions that escape in just one sequence of expansive iterates.

The goal of this section is to find a set $S' \subset X$ that every escaping orbit must visit in order to escape. More precisely, the last time a trajectory lies below an intermediate threshold (specified below) before escaping, its x coordinate must lie in S'. In order to define S', we will consider the x dynamics in symbolic terms. For this, we fix a Markov partition $\mathcal{R}$ for T, as in §2.3.1. Growth in the y term happens when a trajectory spends a long time in the expansive neighborhood of x^*. When a transition from expansive to non-expansive sequence (or vice versa) occurs, there is a contraction as described above. We will represent a point x in X by two sequences of numbers: $\xi_0(x), \xi_1(x), \ldots$, indicating the number of consecutive iterates the x trajectory spends in a Markov rectangle containing the fixed point x^* and $\tilde{\xi}_1(x), \tilde{\xi}_2(x), \ldots$, indicating the number of consecutive iterates the x trajectory spends outside of it. We also let $N_k := \xi_0 + \xi_1 + \cdots + \xi_k$ and $\tilde{N}_k := \tilde{\xi}_1 + \cdots + \tilde{\xi}_k$. All of these numbers can be thought of as random variables on the Borel probability space (X, μ). Our set S' will be defined in terms of consecutive sequences of ξ.

Remark 2.4.4. In the case of the maps $T(x) = mx \pmod 1$, there exists a Markov partition for which the sequence of ξ_j corresponds to a sequence of iid geometric ran-

dom variables on the Borel probability space (X, μ). In this case, calculations can be done directly, using only properties of elementary discrete probability distributions.

In general, the random variables ξ_i are not independent. However, the exponential cluster property (also known as ψ mixing property with exponential decay) used in §2.3.1.3 allows one to show, in our parameter range, that the total probability of escape can be still compared with the probability of escaping through only one long sequence of consecutive expanding iterates. This was estimated in §2.4.1.1.

First, we establish a lower bound in the average bursting time in terms of $p\alpha$, that is of special interest in the case when the multiplicative effect is negligible (small $p\alpha$). Later, in Proposition 2.4.6, we establish a sharper lower bound for the multiplicative case (large $p\alpha$).

We set $\Delta = \|g\|_\infty + \zeta = \mathcal{O}(1)$, so that $q\Delta$ is a global upper bound on the kick. For the Markov partition $\mathcal{R}$, we let R_0 be the rectangle containing x^*, and define $\Delta_0 = \sup_{x \in R_0}\{g(x)\} + \zeta$, so that $q\Delta_0$ bounds the kick on R_0. We recall that the partition $\mathcal{R}$ can be chosen with arbitrarily small radius. Hence, Δ_0 can be made as close to $1 + \zeta$ as desired. Also, let $\Gamma = \sup_{x \in R_0}\{h(x) + \sigma\}$; then Γ can be made arbitrarily close to $1 + \sigma$.

Proposition 2.4.5. Let $l(z) = \frac{z}{e^z - 1}$ and let $\epsilon > 0$. If p, q and s are sufficiently small, α is sufficiently large, and $p\alpha \leq \frac{1}{2} \log \frac{1}{c}$, we have:

$$\log \tau \geq (1 - \epsilon) l(p\alpha) \Lambda \alpha.$$

Proof. To establish this, we let $\tilde{\Delta}_0 = \frac{\Delta_0}{l((1+\sigma)p\alpha)}$. We also fix $B > 0$ as in the statement of Lemma 2.3.5 and choose time 0 to be the last time that the y trajectory is below

Bq, so that for $n \geq 1$, y_n never goes back below Bq before exceeding the threshold Y. Notice that for $x_n \in R_0$,

$$y_{n+1} \leq \big(f(x_n) + (h(x_n) + \sigma)p\big)y_n + q\Delta_0 \leq (1 + \Gamma p)y_n + q\Delta_0.$$

First, we consider only escaping trajectories for which $\xi_i \leq \alpha$ for all i before escaping. The measure of the remaining escaping trajectories will be included directly in the final estimate. Then, by induction on ξ, for $\xi \leq \alpha$,

$$y_\xi - y_0(1 + \Gamma p)^\xi \leq q\Delta_0 \frac{(1 + \Gamma p)^\xi - 1}{\Gamma p} \leq q\Delta_0 \frac{e^{\Gamma p \xi} - 1}{\Gamma p} \leq q\tilde{\Delta}_0 \xi.$$

Next, for $x_n \notin R_0$ we have $y_{n+1} \leq cy_n + q\Delta$, so by induction on $\tilde{\xi}$,

$$y_{\xi+\tilde{\xi}} \leq c^{\tilde{\xi}}(y_0(1 + \Gamma p)^\xi + q\tilde{\Delta}_0 \xi) + \frac{q\Delta}{1 - c}.$$

Let $\tilde{c} = c(1 + \Gamma p)^\alpha$. Then $\tilde{c} < 1$ if p, q and s are sufficiently small and $p\alpha \leq \frac{1}{2} \log \frac{1}{c}$. By induction, we obtain:

$$y_{N_k + \tilde{N}_k} \leq \tilde{c}^k y_0 + \frac{q\Delta}{(1 - c)(1 - \tilde{c})} + q\tilde{\Delta}_0 \Big(\sum_{j=0}^{k} \tilde{c}^{k-j} \xi_j \Big).$$

If the threshold $Y = \alpha q$ is reached within $k > 1$ expansive sequences, then recalling that $y_0 \leq Bq$ we must have:

$$\sum_{j=0}^{k} \tilde{c}^{k-j} \xi_j \geq \frac{\alpha - \tilde{c}^k B}{\tilde{\Delta}_0} - \frac{\Delta}{\tilde{\Delta}_0 (1 - c)(1 - \tilde{c})}.$$

In this context, we will say that a trajectory escapes by route k if $k+1$ is the smallest integer for which the above holds. The set S' mentioned above consists of the union of trajectories that may initiate an escape by route k over all $k \in \mathbb{N}$, and those for which there exists some i with $\xi_i > \alpha$ before escaping. We let ι be the smallest so that $\xi_\iota > \alpha$, and denote its measure by $\hat{\mu}_\iota$.

To bound $\hat{\mu}_\iota$, we use the proof of Lemma 2.3.5 with $t = \alpha$ and $c = \tilde{c}$ defined above. Adding over ι, we get that $\sum_{\iota\in\mathbb{N}} \hat{\mu}_\iota \leq \hat{A}\chi^{-\alpha}$, for some constant $\hat{A}$ independent of p, q and α.

Given any $\epsilon > 0$, if α sufficiently large (depending on ϵ but independent of p and q), we can apply Lemma 2.3.5 with $t = (1-\epsilon)\frac{\alpha}{\tilde{\Delta}_0} > B$ and $t_* = B$. We obtain that the measure μ_k of x trajectories that initiate an escape by route k decays exponentially with k. For $k = 1$, Lemma 2.3.4 implies that $\mu_1 \leq A\chi^{-(1-\epsilon)\frac{\alpha}{\tilde{\Delta}_0}}$.

Combining the previous two paragraphs, we get that there exists some constant $\tilde{L}$ such that the total measure $\mu(S')$ is bounded by

$$\mu(S') \leq \tilde{L}\chi^{-(1-\epsilon)\frac{\alpha}{\tilde{\Delta}_0}}.$$

Recalling that $\sigma \to 0$ as $(p, q, s) \to (0, 0, 0)$ and that Δ_0 can be chosen arbitrarily close to 1 by choosing $\mathcal{R}$ appropriately, and (p, q, s) sufficiently close to $(0, 0, 0)$, we combine the previous estimate with Lemma 2.3.2, and conclude that for α sufficiently large, $p\alpha \leq \frac{1}{2}\log\frac{1}{c}$ and p, q and s sufficiently small we have:

$$\log\tau \geq (1-\epsilon)l(p\alpha)\Lambda\alpha. \qquad \square$$

Recall that $K(p,\alpha) := \frac{1}{p\alpha\tilde{l}(p\alpha)}e^{\frac{p}{2}\chi^{\frac{\log\frac{1}{c}}{4p(1+\tilde{l}(p\alpha))}}}$.

Proposition 2.4.6. Let $\epsilon > 0$. For sufficiently small p, q and s for which $p\alpha \leq \frac{1}{2}\log\frac{1}{c}$ and $K(p,\alpha) > 1$, a lower bound on the scaling of $\log\tau$ is:

$$\log\tau \geq (1-\epsilon)\big(1-\tilde{l}(p\alpha)\big)\Lambda\frac{\log(1+p\alpha)}{p},$$

where $\tilde{l}(p\alpha) \to 0$ as $p\alpha \to \infty$.

Remark 2.4.7. The restriction on the size of α in terms of p can be improved by taking into account the fact that typically, trajectories spend a long time outside of the expanding region before coming back to it. This would allow larger thresholds α. However, sufficiently large values of α would still need to be excluded. Therefore, we only present the argument as stated in Proposition 2.4.6.

Proof of Proposition 2.4.6. Now we fix $B > 0$, to be specified later, and for the moment choose time 0 to be the first time that the y trajectory exceeds Bq and y_n never goes back below Bq before escaping. After a sequence of expansions corresponding to a block of length ξ followed by a contraction, similarly to the proof of Proposition 2.4.5, we have:

$$\begin{aligned}\frac{y_{\xi+\tilde{\xi}}}{y_0} &\leq \left((1+\Gamma p)^{\xi} + \tfrac{\Delta_0}{B}\tfrac{(1+\Gamma p)^{\xi}-1}{\Gamma p}\right)c + \tfrac{\Delta}{B(1-c)} \\ &\leq (1+\Gamma p)^{\xi}\left(1 + \tfrac{\Delta}{B(1-c)c} + \tfrac{\Delta_0}{B}\tfrac{1-(1+\Gamma p)^{-\xi}}{\Gamma p}\right)c \\ &\leq (1+\Gamma p)^{\xi}\left(1 + \tfrac{\Delta}{B(1-c)c} + \tfrac{\Delta_0}{B}\xi\right)c.\end{aligned}$$

Let $E(p) := \log(1+\Gamma p)$. Then, by induction on k,

$$\begin{aligned}\log\left(\frac{y_{N_k+\tilde{N}_k}}{y_0}\right) &\leq N_k E(p) + k\log c + \textstyle\sum_{j=1}^{k}\log\left(1 + \tfrac{\Delta}{B(1-c)c} + \tfrac{\Delta_0}{B}\xi_j\right) \\ &\leq N_k\left(E(p) + \tfrac{\Delta_0}{B}\right) + k\left(\log c + \tfrac{\Delta}{B(1-c)c}\right).\end{aligned}$$

Therefore, for a trajectory to initiate an escape without returning to the region $y \leq Bq$ before reaching the threshold, we need to have the following inequality holding for some k, l:

$$\log\frac{\alpha}{B} \leq (N_{k+l} - N_{l-1})\left(E(p) + \frac{\Delta_0}{B}\right) + k\left(\log c + \frac{\Delta}{B(1-c)c}\right).$$

Equivalently,

$$\begin{aligned} N_{k+l} - N_{l-1} &\geq \frac{\log \frac{\alpha}{B}}{E(p)+\frac{\Delta_0}{B}} - k\frac{\log c + \frac{\Delta}{B(1-c)c}}{E(p)+\frac{\Delta_0}{B}} \\ &=: M_0(\alpha, p, B) + k\beta(p, B) =: M_k(\alpha, p, B). \end{aligned}$$

We will say that such a trajectory escapes by route k. This condition depends only on the x dynamics and will be used to bound the total measure of trajectories that initiate an escape by route k from above. In this setting, we define the set $S' \subset X$ as the union of all trajectories that can initiate an escape by route k over all $k \in \mathbb{N}$.

By Lemma 2.3.6, we know that if β is sufficiently large and M_0 is not exponentially large in β, there is a constant $0 < \tilde{\theta} < 1$ such that

$$\mu(N_{k+l} - N_{l-1} \geq M_0 + k\beta) \leq A\tilde{\theta}^k \chi^{-M_0}.$$

Now, we set $B = \frac{\alpha \log(1+p\alpha)}{1+p\alpha}$, and recall that $\tilde{l}(z) = \frac{1+z}{z\log(1+z)}$. For p, q and s sufficiently small, the restriction in the size of M_0 relative to β is satisfied as long as

$$K(p, \alpha) = \frac{B}{\alpha} e^{\frac{p}{2}\chi^{\frac{\log \frac{1}{c}}{4(p+1/B)}}} > 1.$$

Then, by Lemma 2.3.2, the measure of S' is bounded by $\mu(S') \leq \frac{A}{1-\tilde{\theta}}\chi^{-M_0}$. In consequence, for sufficiently small p, q and s we have:

$$\begin{aligned} \log \tau &\geq \Lambda \frac{\log(1+p\alpha) - \log\log(1+p\alpha)}{\log(1+\Gamma p) + \Delta_0 \frac{1+p\alpha}{\alpha \log(1+p\alpha)}} + \log \frac{A}{1-\tilde{\theta}} - \log 4 \\ &\geq (1-\epsilon)\Lambda \frac{\log(1+p\alpha)}{p(1+\tilde{l}(p\alpha))} \geq (1-\epsilon)(1-\tilde{l}(p\alpha))\Lambda \frac{\log(1+p\alpha)}{p}. \end{aligned}$$

□

If parameters p, α satisfy condition $(\star)$, upper and lower bounds from Propositions 2.4.3 and 2.4.6 combined yield Theorem 2.4.

Remark 2.4.8. In case the bifurcating orbit is periodic, $\{x_1^*, \ldots, x_d^*\}$, the corresponding f and h in the analogue of Equation (2.3) for T^d have the same value at all points of the bifurcating periodic orbit. Furthermore, there are smooth conjugacies between the fiber maps restricted to small neighborhoods of the d fixed points. In general, we would not be able to normalize g simultaneously at all d points; instead, we would normalize g so that its maximum value on the periodic orbit is 1. The estimates for the lower bound would need to be modified accordingly. The ones for the upper bound remain valid.

2.4.2 Proof of scaling laws

In this section, we extend the linear analysis presented in §2.4.1 to the nonlinear setting, and complete the proof of the results stated in §2.2.2. We also obtain results that are valid in a parameter range broader than that of Theorems 2.1 and 2.2, as claimed in the introduction.

With the normalizations described in §2.2.1 and after possibly rescaling y, the y dynamics on the fiber over the fixed point x^* is described as follows. In the case of transcritical bifurcations (general case), Equation (2.4) becomes

$$y_{n+1} = (1+p)y_n \pm y_n^2 + \mathcal{O}(qy_n + p^2 y_n + pq + q^2 + y_n^3), \tag{2.11}$$

and in the case of pitchfork bifurcations (symmetric case), Equation (2.5) becomes

$$y_{n+1} = (1+p)y_n \pm y_n^3 + \mathcal{O}(qy_n + p^2 y_n + pq + q^2 + y_n^4). \tag{2.12}$$

In [ZHO03], Zimin, Hunt and Ott have classified the effect of the nonlinearities depending on whether they accelerate or confine the burst. They call them hard and soft transitions, respectively. We will analyze these two scenarios. We also distinguish between multiplicative (drift-dominated) and additive (noise-dominated) bubbling phenomena, which occur depending on the relative sizes of the parameters p and q. Roughly speaking, when the effect of p is dominant, we call it multiplicative bubbling, and when it is negligible, we call it additive bubbling.

We note that the analysis from §2.4.1.2 is applicable in the nonlinear setting since it deals with a lower bound for the bursting time. On the other hand, we have to adjust the upper bound estimates from §2.4.1.1 to incorporate nonlinear terms.

2.4.2.1 Asymmetric case: generic transcritical bifurcation.

Here, we show two scaling laws valid for generic asymmetric bubbling bifurcations. They are valid for a threshold Y independent of p and q in the hard transition case ($a(x^*)g(x^*) > 0$), proportional to p in the multiplicative case of soft transition ($a(x^*)g(x^*) < 0$), and to $\sqrt{q}$ in the additive case of soft transition, as will be shown in the proofs. In this setting, the y dynamics of the fixed point x^* can be written as (2.11).

Multiplicative bubbling.

Proposition 2.4.9. If $p^2 > 4q > \frac{p}{k(p)}$, there exists a constant $\tilde{C} > 1$ independent of p, q such that if (p, q) is sufficiently close to $(0, 0)$,

$$\tilde{C}^{-1}\Lambda \leq \frac{\log \tau(Y)}{\frac{1}{p}\log(1 + \frac{p^2}{q})} \leq \tilde{C}\Lambda.$$

Furthermore, for any $\epsilon > 0$, if $(p, \frac{q}{p^2})$ is sufficiently close to $(0, 0)$,

$$(1 - \epsilon)\Lambda < \frac{\log \tau(Y)}{\frac{1}{p}\log \frac{p^2}{q}} < (1 + \epsilon)\Lambda.$$

Proof. Assume is $p^2 \geq 4q$. The attracting fixed point of the y dynamics is our threshold of interest in the soft transition case ($a = -1$). Since $y_* \approx p$, we set $\alpha = r\frac{p}{q}$, for some $0 < r \leq 1$. In this case, the parameter s introduced in §2.4.1 is simply $r = s$, so $s \to 0$ if $r \to 0$.

The hard transition case ($a = 1$), where no attractor is given by the local analysis, corresponds to the scenario where a linear regime takes place and then it is replaced by a nonlinear one. We set a threshold $\alpha = r\frac{p}{q}$ to separate the linear and nonlinear behaviors, for some $r > 0$ independent of p and q.

Let us take an initial condition $y_0 = \frac{q}{p}$. Assuming p is small and q is small but not extremely small compared to p, $\frac{1}{p}k(p) > \frac{1}{q}$, Theorem 2.4 implies the following scaling for sufficiently small r:

$$C^{-1}\Lambda < \frac{\log \tau(rp)}{\frac{1}{p}\log(1 + \frac{rp^2}{q})} < C\Lambda,$$

which, in turn, implies:

$$C^{-1}\frac{\log(1 + \frac{rp^2}{q})}{\log(1 + \frac{p^2}{q})}\Lambda < \frac{\log \tau(rp)}{\frac{1}{p}\log(1 + \frac{p^2}{q})} < C\Lambda.$$

In the hard transition case, the burst is not confined in a small region. It may be of order one. In this setting, we also investigate the average bursting time associated to a threshold Y, which is determined by the y value at which the higher order terms become significant, for example of size $\frac{1}{3}y^2$. To bound $\log \tau(Y)$ from below, we use $\tau(Y) \geq \tau(rp)$ for $rp < Y$. We choose a threshold $Y \leq 1$, that is reached for all sufficiently small values of p and q and such that the higher order terms are bounded by $\frac{1}{3}y^2$ for $rp < y < Y$.

To find an upper bound on the scaling of $\log \tau$, we extend the analysis in §2.4.1.1. There we found $\tilde{n}$ such that if x spends $\tilde{n}$ consecutive iterates in the region $|x-x^*| < \tilde{x}$, then at the end of those iterations, $y \geq \alpha q = rp$. We can guarantee that $y \geq Y$ if x spends t additional iterates in the region $|x-x^*| \leq \tilde{x}$, where we determine t as follows. When $|x_n-x^*| < \tilde{x}$ and $rp \leq y_n \leq Y$, $y_n+\frac{2}{3}y_n^2 \leq y_{n+1} \leq (1+2p)y_n+\frac{4}{3}y_n^2$. Hence, we have that $y_{n+1} \leq \frac{8}{3}y_n$.

Calling the time at which y exceeds rp time 0, we can bound from below the solution of our original difference equation with the solution $y(t)$ of a differential equation inductively if we can check $y(0) = rp$ and $y(n+1) \leq y(n) + \frac{2}{3}y(n)^2$. For values $n \leq \frac{32}{3rp} - \frac{32}{3}$, this is the case for the solution of

$$\dot{y} = \frac{3}{32}y^2, \quad y(0) = rp.$$

This solution is given by $y(t) = \frac{1}{\frac{1}{rp}-\frac{3}{32}t}$.

From this, we conclude that an extra $t = \frac{32}{3}\frac{(Y-rp)}{Yrp} \leq \frac{32}{3rp} - \frac{32}{3} < \frac{32}{3rp}$ iterates in the non-contracting region would oblige a burst of size Y. Thus, proceding as in

(2.10), for p, q and r sufficiently small, we have the following bounds:

$$C^{-1}\frac{\log(1+\frac{rp^2}{q})}{\log(1+\frac{p^2}{q})}\Lambda < \frac{\log\tau(Y)}{\frac{1}{p}\log(1+\frac{p^2}{q})} < C\left(1+\frac{1}{r\log(1+\frac{p^2}{q})}\right)\Lambda.$$

In particular, if we fix a sufficiently small value for r, the first statement follows.

We obtain the second statement, corresponding to the asymptotic scaling for $\log\tau(Y)$ in the parameter regime considered in [ZHO03], $p^2 \gg q$ as follows. For any $\epsilon > 0$, if (p, q) is sufficiently close to $(0, 0)$ and $\frac{p^2}{q}$ sufficiently large, we can let $r = \frac{1}{\log\log\frac{p^2}{q}}$ and obtain from the previous bounds:

$$(1-\epsilon)\Lambda < \frac{\log\tau(Y)}{\frac{1}{p}\log\frac{p^2}{q}} < (1+\epsilon)\Lambda. \qquad \square$$

Additive bubbling.

Proposition 2.4.10. If $p^2 < 4q$, there exists a constant $\tilde{C} > 1$ independent of p, q such that if (p, q) is sufficiently close to $(0, 0)$,

$$\tilde{C}^{-1}\Lambda \leq \frac{\log\tau(Y)}{\frac{1}{q^{\frac{1}{2}}}} \leq \tilde{C}\Lambda.$$

Proof. Assume $p^2 < 4q$. In the case of a soft transition ($q > 0$), the attracting fixed point for the y dynamics is $y_* \approx \sqrt{q}$. Our threshold of interest is of the order of $\sqrt{q}$. Hence, we choose $\alpha = r\frac{1}{\sqrt{q}}$. In this case, $r = \sqrt{s}$ and condition s is small when r is small.

In the hard transition case, the linear term is negligible with respect to the kick. Therefore nonlinear terms become significant when the kick becomes negligible, and no intermediate regime is governed by the expansive linear term. In this setting, we investigate the threshold $\alpha q = r\sqrt{q} \approx y_*$, which separates the constant and nonlinear behaviors.

In both cases we first require to reach $\alpha = \frac{r}{\sqrt{q}}$, for some $0 < r \leq 1$ sufficiently small, corresponding to the predominance of the linear regime. From Theorem 2.4, if p, q and r are sufficiently small, we obtain:

$$C^{-1}(1-r)\Lambda < \frac{\log \tau(r\sqrt{q})}{\frac{r}{\sqrt{q}}} \leq C\Lambda.$$

In the hard transition case, by reasoning similarly to the multiplicative case, we obtain that to pass from the linear setting to the threshold Y of order 1, $\frac{32}{3r\sqrt{q}}$ extra iterates in the non-contracting region suffice. Hence, we have:

$$C^{-1}r(1-r)\Lambda \leq \frac{\log \tau(Y)}{\frac{1}{q^{\frac{1}{2}}}} \leq \Lambda\left(Cr + \frac{32}{3r}\right).$$

Hence, if we fix a sufficiently small value for r, Proposition 2.4.10 follows. □

2.4.2.2 Symmetric case: generic pitchfork bifurcation.

Here, we show two scaling laws valid for generic pitchfork bubbling bifurcations. They are valid for a threshold Y independent of p and q in the hard transition case ($a(x^*) > 0$), proportional to $\sqrt{p}$ in the multiplicative case of soft transition ($a(x^*) < 0$), and to $\sqrt[3]{q}$ in the additive case of soft transition, as will be shown in the proofs. In this setting, the y dynamics of the fixed point x^* can be written as (2.12).

Multiplicative bubbling.

Proposition 2.4.11. If $p^3 > \frac{27}{4}q^2$ and $q > \frac{\sqrt{p}}{k(p)}$, there exists a constant $\tilde{C} > 1$ independent of p, q such that if (p, q) is sufficiently close to $(0, 0)$,

$$\tilde{C}^{-1}\Lambda \leq \frac{\log \tau(Y)}{\frac{1}{p}\log(1 + \frac{p^{\frac{3}{2}}}{q})} \leq \tilde{C}\Lambda.$$

Furthermore, for any $\epsilon > 0$, if $(p, \frac{p^{\frac{3}{2}}}{q})$ is sufficiently close to $(0, 0)$,

$$(1 - \epsilon)\Lambda < \frac{\log \tau(Y)}{\frac{1}{p}\log(1 + \frac{p^{\frac{3}{2}}}{q})} < (1 + \epsilon)\Lambda.$$

Proof. Assume $p^3 \geq \frac{27}{4}q^2$. The soft transition case occurs when $a = -1$. In this situation, the cubic equation has three real roots and the continuation of the fixed point 0, $y_* \approx \sqrt{p}$, is stable. In this case, we set $\alpha = r\frac{\sqrt{p}}{q}$, where r corresponds to $\sqrt{s}$, and therefore s is small when r is.

The hard transition case occurs when $a = 1$. The threshold corresponding to $\alpha = r\frac{\sqrt{p}}{q}$ corresponds to the transition between linear and nonlinear behaviors.

The analysis is similar to the previous subsection. Let us take an initial condition $y_0 = \frac{q}{p}$. Assuming that p is small and $\frac{1}{\sqrt{p}}k(p) > \frac{1}{q}$, by Theorem 2.4 we get:

$$C^{-1}\frac{\log(1 + r\frac{p^{\frac{3}{2}}}{q})}{\log(1 + \frac{p^{\frac{3}{2}}}{q})}\Lambda \leq \frac{\log \tau(r\sqrt{p})}{\frac{1}{p}\log(1 + \frac{p^{\frac{3}{2}}}{q})} \leq C\Lambda.$$

As in the asymmetric case, when the transition is hard, we are also interested in bursts up to order one, whose size Y is determined by higher order terms, but independent of p and q. We choose it in such a way that the higher order terms are bounded by $\frac{1}{3}y^3$ for $r\sqrt{p} \leq y < Y$. In this case, if $|x_n - x^*| \leq \tilde{x}$ and $y_n \geq r\sqrt{p}$, we know that $y_n + \frac{2}{3}y_n^3 \leq y_{n+1} \leq (1 + 2p)y_n + \frac{4}{3}y_n^3$.

As in §2.4.2.1, $\frac{y_n}{y_{n+1}} \geq \frac{3}{8}$, and we consider the differential equation:

$$\dot{y} = \frac{9}{256} y^3, \quad y(0) = r\sqrt{p},$$

with solution given by $y(t) = \sqrt{\frac{r^2 p}{1 - 2\frac{9}{256} r^2 p t}}$.

This function bounds from below the solution of our system up to $t = \frac{128}{9r^2 p} - \frac{128}{9}$. This is the time it takes the solution of the differential equation to reach Y. Hence, we get that an extra $t = \frac{128}{9r^2 p}$ iterates in the non-contracting region would oblige a burst of size Y. Thus, if p, q and r are sufficiently small, we have the following bounds:

$$C^{-1} \frac{\log(1 + r\frac{p^{\frac{3}{2}}}{q})}{\log(1 + \frac{p^{\frac{3}{2}}}{q})} \Lambda < \frac{\log \tau(Y)}{\frac{1}{p} \log(1 + \frac{p^{\frac{3}{2}}}{q})} < C \left(1 + \frac{1}{r^2 \log(1 + \frac{p^{\frac{3}{2}}}{q})}\right) \Lambda.$$

Hence, if we fix a sufficiently small value for r, the first statement follows. Furthermore, we obtain the asymptotic scaling for $\log \tau(Y)$ in the parameter regime considered in [ZHO03], $p^{\frac{3}{2}} \gg q$, as follows. For any $\epsilon > 0$, if (p, q) is sufficiently close to $(0, 0)$ and $\frac{p^{\frac{3}{2}}}{q}$ sufficiently large, we can let $r = \frac{1}{\log \log \frac{p^{\frac{3}{2}}}{q}}$ and obtain from the previous bounds:

$$(1 - \epsilon)\Lambda < \frac{\log \tau(Y)}{\frac{1}{p} \log(1 + \frac{p^{\frac{3}{2}}}{q})} < (1 + \epsilon)\Lambda. \qquad \square$$

Additive bubbling.

Proposition 2.4.12. If $p^3 < \frac{27}{4} q^2$, there exists a constant $\tilde{C} > 1$ independent of p, q such that if (p, q) is sufficiently close to $(0, 0)$,

$$\tilde{C}^{-1} \Lambda \leq \frac{\log \tau(Y)}{\frac{1}{q^{\frac{2}{3}}}} \leq \tilde{C} \Lambda.$$

Proof. Assume $p^3 < \frac{27}{4}q^2$. Analogously to the asymmetric case, we first consider the linear regime, determined by the fact that $y_* \approx \sqrt[3]{q}$. We set the threshold $\alpha q = r\sqrt[3]{q}$. In this setting, $r = \sqrt[3]{s}$, and from Theorem 2.4, for $\epsilon > 0$, if p, q and r are sufficiently small, we obtain:

$$C^{-1}(1-r)\Lambda \leq \frac{\log \tau(r\sqrt[3]{q})}{\frac{r}{q^{\frac{2}{3}}}} \leq C\Lambda.$$

As above, in the hard transition case, to pass from the linear setting to a threshold Y of order 1, $\frac{128}{9r^2q^{\frac{2}{3}}}$ extra iterates in the non-contracting region suffice. Hence, for sufficiently small p, q and r we have:

$$C^{-1}r(1-\frac{3}{2}r)\Lambda \leq \frac{\log \tau(Y)}{\frac{1}{q^{\frac{2}{3}}}} \leq \left(Cr + \frac{128}{9r^2}\right)\Lambda.$$

Hence, if we fix a sufficiently small value for r, Proposition 2.4.12 follows. □

2.5 Random mismatch for symmetric systems

In this section, we carry out the analysis in the case when the perturbation $qg(x_n)$ in (2.3) is replaced by an additive noise term of the form qg_n, where $\{g_n\}_{n\in\mathbb{N}}$ is a sequence of independent random variables identically distributed on $[-1, 1]$, according to a probability distribution $\mathbb{P}$. This could represent a random noise or mismatch in a synchronized system. We choose q to be the maximum noise amplitude, so that -1 and/or 1 is in contained in the support of $\mathbb{P}$.

In this case, we can treat bursts in the negative y direction in the same way as bursts in the positive y direction. In the asymmetric case, having bursts in both directions implies a hard transition, but the possibility that bursts may be

more likely to initiate in the opposite direction from the y^2 term makes the average bursting time harder to bound from above.

In this section, we treat the case of symmetric y dynamics. A significant part of our previous analysis remains applicable in both the symmetric and asymmetric cases. In particular, since q bounds the noise term, the lower bound on the average bursting time remains unchanged. For the upper bound, we have to find the expected time to get a long sequence of *coherent* kicks, enough to push the trajectory beyond the threshold. By a coherent sequence, we mean that there exists some $\tilde{\delta} > 0$ such that $|g_k| > \tilde{\delta}$ for a sequence of consecutive values of k and all these g_k have the same sign. In the deterministic case, we were able to choose some $\tilde{\delta} > 0$ so that whenever a trajectory was sufficiently close to x^*, the kick was of size at least $\tilde{\delta}$. In the random case, we will choose $\tilde{\delta}$ depending on the distribution of the noise. In what follows, we assume $\mathbb{P}$ has negative and positive values in its support. If this was not the case, the analysis from the deterministic case would be applicable, with $\tilde{\delta}$ a lower bound on g_k and the upper bound on the scaling from Proposition 2.4.3 changed to

$$\frac{\log \tau}{\frac{1}{p}\log(1+\frac{p\alpha}{\tilde{\delta}})} < (1+\epsilon)\Lambda.$$

Remark 2.5.1. Our methods are also suitable to study the case of combined random noise and deterministic perturbation. The case that the support of $\mathbb{P}$ has both positive and negative values corresponds to the case where the size of the noise is larger than the perturbation. Though our analysis here only treats the case where the perturbation is independent of x, it is not hard to remove this restriction.

Multiplicative case.

We proceed as in Section 2.4.1.1 to establish an upper bound for the bursting time. As in the proof of Proposition 2.4.3, we can find $\tilde{x} = \mathcal{O}(p)$ so that for $|x - x^*| < \tilde{x}$ we have

$$f(x) + (h(x) - \sigma)p > 1 + \tilde{\gamma}p \quad \text{for some } 0 < \tilde{\gamma} < 1 - \sigma, \text{ and also}$$

$$\mathbb{P}(g > \tilde{\delta}) > 0 \text{ and } \mathbb{P}(g < -\tilde{\delta}) > 0 \text{ for some } \tilde{\delta} > 0.$$

As in Section 2.2.1, the y dynamics are given by (2.5). Due to the symmetry in y, a sufficiently long sequence of kicks in the same direction combined with expansion, guarantee a soft burst, which in the multiplicative case corresponds to $y \approx \sqrt{p}$. As in Section 2.4.2.2, an extra number of expansive iterates implies a hard burst. Hence, for sufficiently small values of p and q, we have in the hard transition case $a(x^*) > 0$:

Proposition 2.5.2. If $p^3 \geq \frac{27}{4}q^2$ and $q > \frac{\sqrt{p}}{k(p)}$, p and q are sufficiently small, there exist a constant $C > 1$ and a threshold Y independent of p, q such that

$$C^{-1}\Lambda \leq \frac{\log \tau(Y)}{\frac{1}{p}\log(1 + \frac{p^{\frac{3}{2}}}{q})} \leq C\frac{\log(1 + \frac{p^{\frac{3}{2}}}{\tilde{\delta}q})}{\log(1 + \frac{p^{\frac{3}{2}}}{q})}(\Lambda - \log\min\{\mathbb{P}(g > \tilde{\delta}), \mathbb{P}(g < -\tilde{\delta})\}).$$

The last term comes from the requirement of a long sequence of *coherent* noise. This is essential, since in the case of non-coherent realizations of the noise, trajectories can spend several steps close to $|y| = \frac{q}{p}$. This does not happen for the deterministic case due to the instability of the fixed point of (2.5) which is close to $\frac{q}{p}$.

In the particular case of noise given by Bernoulli random variables with parameter $\frac{1}{2}$, under the same assumptions on parameters as in the above proposition, the scaling law is simplified to:

$$C^{-1}\Lambda \leq \frac{\log \tau(Y)}{\frac{1}{p}\log(1+\frac{p^{\frac{3}{2}}}{q})} \leq C(\Lambda + \log 2).$$

Additive case.

Assume p, q and the non-linearity parameter s are sufficiently small. The following upper bound for the bursting time is obtained by considering the escaping route given by the occurrence of sufficient consecutive non-contracting coherent kicks. In this setting, we choose $\tilde{\delta}$ such that

$$\mathbb{P}(g > 2\tilde{\delta}) > 0, \quad \mathbb{P}(g < -2\tilde{\delta}) > 0 \text{ for some } \tilde{\delta} > 0 \text{ and}$$

$$f(x) + h(x)p > 1 - \tfrac{\tilde{\delta}}{\alpha},$$

so estimates in the additive case of Section 2.4.1.1 apply.

The upper bound from Section 2.4.1.1 corresponds to $\tilde{\tilde{n}} = \frac{\alpha}{\tilde{\delta}}$ non-contracting iterates, and the additional term in the estimate below comes from the requirement of $\tilde{\tilde{n}}$ consecutive *coherent* kicks, which was automatic in the deterministic case:

$$\log \tau(\alpha q) \leq C(\Lambda - \log \min\{\mathbb{P}(g > 2\tilde{\delta}), \mathbb{P}(g < -2\tilde{\delta})\})\alpha,$$

for some $C > 1$ independent of p, q.

Furthermore, when the drift is negligible, a number of the order of α of coherent kicks are needed to escape. Using arguments analogous to Section 2.4.1.2 we can

find a constant $C > 1$ such that:

$$C^{-1}\Lambda\alpha \leq \log \tau(\alpha q).$$

Combining the two estimates and assuming p and q are sufficiently small yields:

Proposition 2.5.3. For $p^3 \leq \frac{27}{4}q^2$, there exists a constant $C > 1$ independent of p, q such that

$$C^{-1} \log \Lambda \leq \frac{\log \tau(\alpha q)}{\alpha} \leq C(\Lambda - \log \min\{\mathbb{P}(g > 2\tilde{\delta}), \mathbb{P}(g < -2\tilde{\delta})\}).$$

Chapter 3

Approximating invariant densities of metastable systems

3.1 Introduction

Metastable systems are studied in relation with phenomena ranging from molecular [MDHS06] to oceanic [FPET07] dynamics. Typical trajectories of these systems remain in one of its almost invariant (metastable or quasi-stationary) components for a relatively long period of time, but eventually switch to a different component and repeat this behavior. Quantitative aspects of these phenomena have been studied through eigenvalue and eigenvector approximation techniques for Markov models [MSF05, FP08]. Here, we are concerned with rigorous approximation results for eigenvectors–in particular those that correspond to stationary measures of the dynamics–in a more general (non-Markov) setting.

Broadly, our setting concerns the approximation of absolutely continuous invariant probability measures (ACIMs) for certain hyperbolic maps with metastable states. These systems arise from perturbing an initial system T_0 with two disjoint invariant sets I_l, I_r of positive Lebesgue measure. The initial map has two mutually singular ergodic ACIMs, μ_l and μ_r. When T_0 is perturbed in such a way that I_l and I_r lose their invariance and the perturbed map T_ϵ has only one ACIM μ_ϵ, we are interested in approximating μ_ϵ using μ_l and μ_r. Specifically, the systems we consider are piecewise C^2 expanding maps of an interval; see Figure 3.1.

Our results can be understood in the context of dynamical systems with holes as follows. As the invariance of the two initially invariant sets is destroyed by the perturbation, we think of the small set of points $I_l \cap T_\epsilon^{-1} I_r$ that switch from I_l to I_r, and likewise the set $I_r \cap T_\epsilon^{-1} I_l$, as being holes in the initially invariant sets. From this point of view we expect to be able to approximate μ_ϵ, for small ϵ, by a convex combination $\alpha\mu_l + (1-\alpha)\mu_r$ of the two initially invariant measures, with the ratio $\alpha/(1-\alpha)$ depending on the relative sizes of the holes.

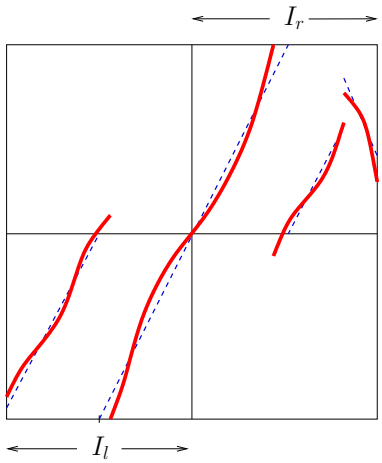

Figure 3.1: Dashed: initial system. Thick: metastable system.

Before discussing our results, we present two illustrative examples. We begin with a simple random system. Consider the family of Markov chains in two states l and r, with transition matrices

$$Q_\epsilon = \begin{pmatrix} 1-\epsilon_{l\to r} & \epsilon_{l\to r} \\ \epsilon_{r\to l} & 1-\epsilon_{r\to l} \end{pmatrix},$$

where $\epsilon = (\epsilon_{l\to r}, \epsilon_{r\to l})$. We are interested in the behavior when $\epsilon \approx 0$. When $\epsilon = 0$, the two sets $I_l = \{l\}$ and $I_r = \{r\}$ are invariant, giving rise to the two ergodic

stationary probability measures $\mu_l = \delta_l$ and $\mu_r = \delta_r$. When $\epsilon_{l\to r} > 0$, there is a unique stationary probability measure

$$\mu_\epsilon = \alpha_\epsilon \mu_l + (1-\alpha_\epsilon)\mu_r, \text{ where } \frac{\alpha_\epsilon}{1-\alpha_\epsilon} = \frac{\epsilon_{r\to l}}{\epsilon_{l\to r}}.$$

Observe that the ratio of the weights $\alpha_\epsilon/(1-\alpha_\epsilon)$, i.e. $\mu_\epsilon(I_l)/\mu_\epsilon(I_r)$, is equal to the inverse ratio of the sizes of the holes, $\epsilon_{r\to l}/\epsilon_{l\to r}$.

Next, we consider two billiard tables $\mathcal{D}_l, \mathcal{D}_r$ in the plane, as indicated in Figure 3.2. For $* \in \{l, r\}$, let $T_* : I_* \circlearrowleft$ be the corresponding billiard map, i.e. the Poincaré map for the first return of the billiard flow to $\partial\mathcal{D}_*$. We use $|\partial\mathcal{D}_*|$ to denote the perimeter of $\mathcal{D}_*$. A general reference for hyperbolic billiards is [CM06], where one can find the background for the assertions below. We use the usual coordinates (s, φ) on I_*, where s is arc length on $\partial\mathcal{D}_*$, and $\varphi \in [-\pi/2, +\pi/2]$ is the angle between the outgoing velocity vector and the inward pointing normal vector to $\partial\mathcal{D}_*$. Then it is well known that T_* leaves (normalized) Liouville measure μ_* invariant, where μ_* has the density $\phi_* := d\mu_*/ds\, d\varphi = [2\,|\partial\mathcal{D}_*|]^{-1}\cos\varphi$. Next, for $\epsilon > 0$, let h_ϵ be a subsegment of $\partial\mathcal{D}_l \cap \partial\mathcal{D}_r$ of length ϵ, and let $\mathcal{D}_\epsilon$ be the billiard table resulting after h_ϵ is removed. The corresponding density for the invariant Liouville measure of the billiard map is $\phi_\epsilon = [2(|\partial\mathcal{D}_l| + |\partial\mathcal{D}_r| - 2\epsilon)]^{-1}\cos\varphi$. Thus as $\epsilon \to 0$,

$$\phi_\epsilon \to \alpha\phi_l + (1-\alpha)\phi_r, \text{ where } \frac{\alpha}{1-\alpha} = \frac{|\partial\mathcal{D}_l|}{|\partial\mathcal{D}_r|},$$

provided some care is taken to define all of the density functions involved on the same space. Note that if we define the holes $H_{*,\epsilon} := T_*^{-1}(h_\epsilon \times [-\pi/2, +\pi/2])$, then we can rewrite $\alpha/(1-\alpha) = \mu_r(H_{r,\epsilon})/\mu_l(H_{l,\epsilon})$, so that again the ratio of the weights equals the inverse ratio of the sizes of the holes. This example is most meaningful

when T_l, T_r, and T_ϵ are all ergodic, which is the case for the tables in Figure 3.2; see §8.15 in [CM06].

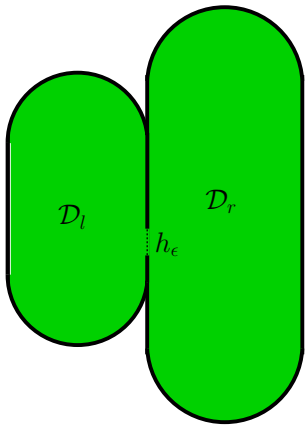

Figure 3.2: Two ergodic billiard tables connected by a hole.

In our main result, Theorem 3.1, we extend the principle behind the examples above to the deterministic setting of piecewise C^2 expanding maps, under fairly general conditions described in §3.2. We show that as $\epsilon \to 0$ the invariant density ϕ_ϵ of T_ϵ converges in L^1 to a convex combination of the ergodic invariant densities of T_0, with the ratio of the weights given by the limiting inverse ratio of the sizes of the holes. We emphasize that our results do not require any of the piecewise expanding maps involved to have a Markov partition.

The density ϕ_ϵ corresponds to an eigenvector with eigenvalue 1 for the Perron-Frobenius operator acting on a suitable space of functions. Our assumptions imply that for $\epsilon > 0$, the operator has 1 as a simple eigenvalue, and also another real simple eigenvalue slightly less than 1. In Theorem 3.2, we characterize the eigenvectors of this lesser eigenvalue by showing that asymptotically they lie on the line spanned

by $d\mu_l/dx - d\mu_r/dx$.

Unlike the two examples above, in the setting of piecewise C^2 expanding maps we have no explicit formulas for the invariant densities; even their existence is non-trivial. Our methods rely on the fact that the densities of the ACIMs for T_ϵ are of bounded variation [LY73]. Hence, they can be decomposed into regular and singular (or saltus) parts, as in [Bal07]. The key technical portions of our proofs include estimating and exploiting the locations and sizes of the jumps at the discontinuities of the invariant densities, which occur on the forward trajectories of the critical points of T_ϵ.

This work is related to other recent work involving metastable systems and piecewise expanding maps. Recently, [KL09] studied metastable systems arising from piecewise smooth uniformly expanding maps with two invariant intervals. They perturbed such an initial map by a family of Markov operators close to the identity to produce a family of metastable systems for $\epsilon > 0$. The associated Perron-Frobenius operators acting on a suitable space of functions have 1 as a simple eigenvalue and another simple eigenvalue $\rho_\epsilon < 1$. As $\epsilon \to 0$, $\rho_\epsilon \to 1$, and the authors rigorously computed the derivative $\lim_{\epsilon \to 0^+}(1 - \rho_\epsilon)/\epsilon$. This provides information on the stationary exchange rate between the metastable states. Their work may be used to show a corresponding result in our setting.

Our work is also related to current and ongoing investigations on linear response. These problems have the feature that, as Ruelle [Rue98] puts it, it is possible to formulate conjectures based on intuition or formal calculations, but the proofs often involve overcoming intricate technicalities. In our setting, we know

that $\mu_\epsilon(I_l) \to \alpha$ as $\epsilon \to 0$. A pertinent open problem would be to try and characterize the higher-order terms $\mathcal{R}(\epsilon) := \mu_\epsilon(I_l) - \alpha$. We do not expect $\mathcal{R}(\epsilon)$ to be differentiable at $\epsilon = 0$ in general. As shown in [BS08], linear response fails precisely when the perturbations T_ϵ are transverse to the topological class of T_0, at least for certain piecewise expanding unimodal maps T_0 that are topologically mixing. Results of [Kel82] show that in that setting the unique ACIM ϕ_ϵ of T_ϵ satisfies $|\phi_\epsilon - \phi_0|_{L^1} = O(\epsilon \log \epsilon)$, where ϕ_0 is the unique ACIM of T_0. [Bal07] gives examples where this estimate is optimal.

Theorem 3.1 can also be regarded as a statement of stochastic stability in the sense that as the size of the perturbation goes to zero, the invariant measures that describe the statistics of Lebesgue almost every orbit have a computable limit. In the uniformly hyperbolic setting, this phenomenon has been studied in the ergodic case; see for example [Via97]. In recent years, there has been work in studying stochastic stability outside the setting of uniformly hyperbolic systems. For example, in [AT05], the authors work at the boundary of expanding maps, in [AAV07] in the context of non-uniformly hyperbolic diffeomorphisms and [Vás07] treats diffeomorphisms with dominated splitting.

Another interesting problem for further research is to extend our results to higher dimensional piecewise hyperbolic maps. While we use techniques specific to one-dimensional maps, we are optimistic that the main elements of our proof, found in §3.3.2, can be generalized.

3.2 Statement of results

In this section, we define a class of dynamical systems with two nearly invariant (metastable) subsets. They are perturbations of a one-dimensional piecewise smooth expanding map with exactly two invariant subintervals I_l and I_r of positive Lebesgue measure. On each of these intervals, the unperturbed system has a unique ACIM. The perturbations break this invariance by introducing what we consider to be holes in the intervals; the hole(s) in I_l map to I_r and vice versa. Each perturbed system will have only one ACIM, and we will determine an asymptotic formula for its density in terms of the invariant densities of the unperturbed system.

Let $I = [0, 1]$. In this paper, a map $T : I \circlearrowleft$ is called a piecewise C^2 map with $\mathcal{C} = \{0 = c_0 < c_1 < \cdots < c_d = 1\}$ as a critical set if for each i, $T|_{(c_i, c_{i+1})}$ extends to a C^2 function on a neighborhood of $[c_i, c_{i+1}]$. We call T uniformly expanding if its minimum expansion, $\inf_{x \in I \setminus \mathcal{C}_0} |T_0'(x)|$, is greater than 1. As is customary for piecewise smooth maps, we consider T to be bi-valued at points $c_i \in \mathcal{C}$ where it is discontinuous. In such cases we let $T(c_i)$ be both values obtained as x approaches c_i from either side, and $T(c_{i\pm})$ the corresponding right and left limits. If $a, b \in \mathcal{C}$, $T|_{[a,b]}$ will be used to specifically denote the restriction of T with $T|_{[a,b]}(a) = T(a_+)$ and $T|_{[a,b]}(b) = T(b_-)$.

We use Leb to denote normalized Lebesgue measure on I and L^1 to denote the space of Lebesgue integrable functions on I, with norm $|f|_{L^1} = \int_I |f(x)|\, dx$. Also, for $f : I \to \mathbb{C}$, we let $|f|_\infty$ be the supremum of f over I and $\mathrm{var}(f)$ be the total

variation of f over I; that is,

$$\mathrm{var}(f) = \sup\{\sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| : n \geq 1, 0 \leq x_0 < x_1 < \cdots < x_n \leq 1\}.$$

For clarity of presentation, we do not state our results under the broadest possible assumptions. However, see §3.2.4 for a number of relaxations of the hypotheses below.

3.2.1 The initial system and its perturbations

We assume that the unperturbed system is a piecewise C^2 uniformly expanding map $T_0 : I \circlearrowleft$ with $\mathcal{C}_0 = \{0 = c_{0,0} < c_{1,0} < \cdots < c_{d,0} = 1\}$ as a critical set. There is a boundary point $b \in (0,1)$ such that $I_l := [0, b]$ and $I_r := [b, 1]$ are invariant under T_0, i.e. for $* \in \{l, r\}$, $T_0|_{I_*}(I_*) \subset I_*$. The existence of an ACIM of bounded variation for $T_0|_{I_*}$ is guaranteed by [LY73]. We assume in addition:

(I1) *Unique ACIMs on the initially invariant set.*

$T_0|_{I_*}$ has only one ACIM μ_*, whose density is denoted by $\phi_* := d\mu_*/dx$.

The uniqueness of such an ACIM can be guaranteed by transitivity or by additional conditions described in [LY78]. From (I1), it follows that all ACIMs of T_0 are convex combinations of the ergodic ones, μ_l and μ_r.

We define the points in $H_0 := T_0^{-1}\{b\} \setminus \{b\}$ to be *infinitesimal holes*. These are all points that map to the boundary point b, except possibly b itself. Our reasons for excluding b from the set of infinitesimal holes will be explained in §3.2.4. An immediate consequence of this definition is that $H_0 \subset \mathcal{C}_0$.

(I2) *No return of the critical set to the infinitesimal holes.*

For every $k > 0$, $(T_0^k \mathcal{C}_0) \cap H_0 = \emptyset$.

This is a non-degeneracy condition that may be difficult to check for specific systems. However, one can show that any piecewise C^2 expanding map of the interval can be approximated by maps satisfying (I2), by making arbitrarily small C^2 perturbations. Such perturbations can be constructed inductively, first adding to T_0 an arbitrarily small C^2 perturbation to obtain $T_{0,1}$ such that $(T_{0,1}\mathcal{C}_0) \cap H_0 = \emptyset$. Successive perturbations should be made so that after the $k-$th perturbation the resulting map $T_{0,k}$ satisfies $(T_{0,k}^k \mathcal{C}_0) \cap H_0 = \emptyset$. Furthermore, each perturbation can be made small enough compared to previous perturbations to guarantee that the sum of the perturbations converges in C^2 and that for each j, the distance between $T_{0,k}^j \mathcal{C}_0$ and H_0 does not decay to zero as $k \to \infty$.

In general, functions of bounded variation are only defined modulo a countable set. However, as we will see in §3.4.2, condition (I2) implies that ϕ_* can be defined so that it is continuous at each of the infinitesimal holes in I_*. Thus it is meaningful to discuss the values of ϕ_* at such points.

(I3) *Positive ACIMs at infinitesimal holes.*

ϕ_l is positive at each of the points in $H_0 \cap I_l$, and ϕ_r is positive at each of the points in $H_0 \cap I_r$.

For example, this will be the case if $T_0|_{I_l}$ and $T_0|_{I_r}$ are weakly covering,[1] see [Liv95].

[1] A piecewise expanding map $T : I \circlearrowleft$ with $\mathcal{C} = \{0 = c_0 < c_1 < \cdots < c_d = 1\}$ as a critical set is weakly covering if there is some N such that for every i, $\cup_{k=0}^{N} T^k([c_i, c_{i+1}]) = I$

(I4) *Restriction on periodic critical points.*

Either

(I4a) $\inf_{x\in I\setminus\mathcal{C}_0}|T_0'(x)|>2$, or

(I4b) T_0 has no periodic critical points, except possibly that 0 or 1 may be fixed points.

Because T_0 may be bi-valued at points in $\mathcal{C}_0$, a critical point $c_{i,0}$ is considered periodic if there exists $n>0$ such that $c_{i,0}\in T_0^n(c_{i,0})$. Condition (I4) is necessary in order to ensure that the perturbed systems defined below satisfy uniform Lasota-Yorke estimates. Since we cannot exclude the possibility of the forward orbit of a critical point containing other critical points, these uniform estimates do not follow directly from the original paper [LY73], but rather from later works, see §3.4.2.

For what follows, we consider C^2-small perturbations $T_\epsilon : I \circlearrowleft$ of T_0 for $\epsilon>0$. This means that a critical set for T_ϵ may be chosen as $\mathcal{C}_\epsilon=\{0=c_{0,\epsilon}<c_{1,\epsilon}<\cdots<c_{d,\epsilon}=1\}$, where for each i, $\epsilon\mapsto c_{i,\epsilon}$ is a C^2 function for $\epsilon\geq 0$. Furthermore, there exists $\delta>0$ such that for all sufficiently small ϵ, there exists a C^2 extension $\hat{T}_{i,\epsilon}:[c_{i,0}-\delta,c_{i+1,0}+\delta]\to\mathbb{R}$ of $T_\epsilon|_{[c_{i,\epsilon},c_{i+1,\epsilon}]}$, and $\hat{T}_{i,\epsilon}\to\hat{T}_{i,0}$ in the C^2 topology. We also assume:

(P1) *Unique ACIM.*

For $\epsilon>0$, T_ϵ has only one ACIM μ_ϵ, with density $\phi_\epsilon:=d\mu_\epsilon/dx$.

(P2) *Boundary condition.*

The boundary point does not move, and no holes are created near the boundary; precisely,

(P2a) If $b \notin \mathcal{C}_0$, then necessarily $T_0(b) = b$. We assume further that for all $\epsilon > 0$, $T_\epsilon(b) = b$.

(P2b) If $b \in \mathcal{C}_0$, we assume that $T_0(b_-) < b < T_0(b_+)$, and also that $b \in \mathcal{C}_\epsilon$ for all ϵ.

If the boundary point does move under the perturbation, condition (P2) often can be satisfied by performing a smooth change of coordinates close to the identity; see §3.2.4.

3.2.2 Main results

The central question of this study is, for small ϵ, how can we asymptotically approximate μ_ϵ by a convex combination of μ_l and μ_r? To that end, let $H_{l,\epsilon} := I_l \cap T_\epsilon^{-1}(I_r)$ and $H_{r,\epsilon} := I_r \cap T_\epsilon^{-1}(I_l)$. We refer to these sets as *holes*. Once a T_ϵ-orbit enters a hole, it leaves one of the invariant sets for T_0 and continues in the other. As $\epsilon \to 0$, the holes converge (in the Hausdorff metric) to the infinitesimal holes from which they arise.

Condition (P1) ensures that for $\epsilon > 0$, at least one of the holes has positive Lebesgue measure. In view of (I3), without loss of generality, we suppose that $\mu_l(H_{l,\epsilon}) > 0$ and define

$$l.h.r. = \lim_{\epsilon \to 0} \frac{\mu_r(H_{r,\epsilon})}{\mu_l(H_{l,\epsilon})},$$

if the limit exists. (*l.h.r.* stands for *limiting hole ratio.*)

Theorem 3.1 (**Approximation of the invariant density**)**.** Consider the family of perturbations T_ϵ of T_0 under the assumptions stated in §3.2.1. Suppose that *l.h.r.*

above exists. Then as $\epsilon \to 0$,

$$\phi_\epsilon \xrightarrow{L^1} \alpha\phi_l + (1-\alpha)\phi_r, \qquad \text{where} \quad \frac{\alpha}{1-\alpha} = l.h.r..$$

We allow for $l.h.r. = +\infty$, in which case $\alpha = 1$. Several straightforward generalizations of the above result are discussed in §3.2.4.

Remark 3.2.1. The limit $l.h.r.$ will always exist as long as the perturbations open up holes $H_{l,\epsilon}$ whose size is truly first order in ϵ: For simplicity, suppose that there are only two infinitesimal holes, $h_l \in I_l$ and $h_r \in I_r$. Then we can always write $H_{*,\epsilon} = (h_* - a_*\epsilon + o(\epsilon), h_* + b_*\epsilon + o(\epsilon))$ for $* \in \{l, r\}$, and if $a_l + b_l > 0$, then

$$l.h.r. = \frac{\phi_r(h_r)(a_r + b_r)}{\phi_l(h_l)(a_l + b_l)}.$$

For example, this will be the case if $T_\epsilon = T_0 + \epsilon g + o(\epsilon)$ for some smooth function g with $g(h_l) > 0$.

The case when the limit $l.h.r.$ does not exist is addressed in §3.2.4.

Remark 3.2.2. An alternative definition of $l.h.r.$ is as a limit of a ratio of escape rates: For $* \in \{l, r\}$, we can consider a dynamical system with a hole, where orbits stop upon entering the hole, by using the unperturbed map $T_0|_{I_*}$ with $H_{*,\epsilon}$ as the hole. See [DY06] for an exposition of such systems. Let $R_{*,\epsilon}$ be the exponential escape rate of Lebesgue measure and suppose that there is only one infinitesimal hole in each initially invariant interval. Then as $\epsilon \to 0$, $\mu_*(H_{*,\epsilon})/R_{*,\epsilon} \to 1$. [BY08, KL09]

Next, let $\mathcal{L}_\epsilon$ be the Perron-Frobenius operator associated with T_ϵ acting on the Banach space $BV = \{f : I \to \mathbb{C} : \mathrm{var}(f) < \infty\}$[2] with the variation norm, and

[2]In fact, we work in the quotient space obtained by identifying two functions of bounded

let $\sigma(\mathcal{L}_\epsilon)$ denote the spectrum of $\mathcal{L}_\epsilon$. It follows from e.g. [Kel89, Thm. 8.3(b)] that $\mathcal{L}_0$ has one as an isolated eigenvalue of multiplicity two. Furthermore in [KL99] the authors show that for fixed small $\delta > 0$ and for every $\epsilon > 0$ small enough, $\sigma(\mathcal{L}_\epsilon) \cap B_\delta(1)$ consists of exactly two eigenvalues, 1 and $\rho_\epsilon < 1$, each of multiplicity 1. As $\epsilon \to 0$, $\rho_\epsilon \to 1$ and the total spectral projection of $\mathcal{L}_\epsilon$ associated with $\sigma(\mathcal{L}_\epsilon) \cap B_\delta(1)$ converges (at a given rate in an appropriate norm) to the total spectral projection of $\mathcal{L}_0$ associated with $\sigma(\mathcal{L}_0) \cap B_\delta(1)$. Note that in §3.4.2 we will verify that a uniform Lasota-Yorke inequality (3.7) holds in our setting. This is the only assumption of [KL99] that is neither trivial nor well-known in our context.

Theorem 3.2 (**Characterization of the eigenspace corresponding to the lesser eigenvalue**)**.** For each $\epsilon > 0$ small enough, there is a unique real-valued function $\psi_\epsilon \in BV$ satisfying $\mathcal{L}_\epsilon \psi_\epsilon = \rho_\epsilon \psi_\epsilon$, $|\psi_\epsilon|_{L^1} = 1$, and $\int_{I_l} \psi_\epsilon dx > 0$. As $\epsilon \to 0$,

$$\psi_\epsilon \xrightarrow{L^1} \frac{1}{2}\phi_l - \frac{1}{2}\phi_r.$$

Remark 3.2.3. Suppose μ_l and μ_r are both mixing for T_0. Given a typical initial density $f \in BV$ (i.e. one with nonzero coefficient of ψ_ϵ when expressed as a linear combination of eigenvectors), as $\mathcal{L}_\epsilon^n f \to \phi_\epsilon$, the deviation $\mathcal{L}_\epsilon^n f - \phi_\epsilon$ becomes approximately proportional to ψ_ϵ for n large.

variation if they differ on at most a countable set. As no confusion arises, we use the same notation for a function and its equivalence class.

3.2.3 Examples

The three piecewise linear maps shown in Figure 3.3 satisfy assumptions (I1)-(I4) from §3.2.1. In all three cases, normalized Lebesgue measure restricted to the left or right intervals is the unique ACIM of the corresponding restricted system.

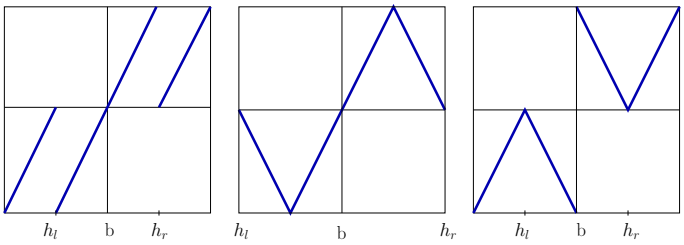

Figure 3.3: Piecewise linear maps giving rise to metastable systems.

Adding a small C^2 perturbation $g : I \times [0, \epsilon_0) \to I$ such that $g(\cdot, 0) \equiv 0$ and for $\epsilon \neq 0$, $g(b, \epsilon) = 0$, $g(h_l, \epsilon) > 0$ and $g(h_r, \epsilon) < 0$ gives a one-parameter family of perturbations $T_\epsilon := T_0 + g(\cdot, \epsilon)$ satisfying assumptions (P1) and (P2).

If $\lim_{\epsilon \to 0} \frac{\text{Leb}(H_{r,\epsilon})}{\text{Leb}(H_{l,\epsilon})} = l.h.r.$, by Theorem 3.1, the invariant densities ϕ_ϵ associated to T_ϵ satisfy

$$\phi_\epsilon \xrightarrow{L^1} \alpha \text{Leb}|_{I_l} + (1-\alpha)\text{Leb}|_{I_r}, \text{ where } \frac{\alpha}{1-\alpha} = l.h.r..$$

The possibility $l.h.r. = \infty$ is allowed, and in this case,

$$\phi_\epsilon \xrightarrow{L^1} \text{Leb}|_{I_l}.$$

Other initial maps T_0 for which Theorems 3.1 and 3.2 are applicable are shown in Figure 3.4.

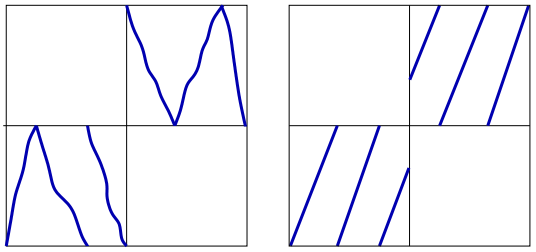

Figure 3.4: Examples of initial maps T_0 which give rise to metastable systems for which our results hold.

3.2.4 Generalizations

Our results extend, with essentially the same proofs, to yield the following straightforward generalizations.

Multiple invariant sets.

We can also allow T_0 to have $m \geq 2$ invariant sets of positive Lebesgue measure $I_1, \ldots, I_m$, provided it has a unique ACIM $\phi_i = d\mu_i/dx$ on each I_i. The invariant sets may be intervals or a union of intervals. See Figure 3.5. For simplicity, we limit ourselves to the case when, for $\epsilon > 0$, all of the transitions between the initially invariant sets are first order in ϵ, i.e. for $i \neq j$, $\mu_i(I_i \cap T_\epsilon^{-1} I_j)$ is either identically 0 or equals $\epsilon \cdot \beta_{i,j} + o(\epsilon)$ for some $\beta_{i,j} > 0$. In this case, under assumptions that are straightforward generalizations of (I1)-(I4), (P1) and (P2), the unique invariant density ϕ_ϵ of T_ϵ converges as $\epsilon \to 0$ to a convex combination of the ϕ_i. The coefficients may be determined from the corresponding coefficients of the stationary measure for the continuous time finite state Markov chain whose transition matrix has the off-diagonal entries $\epsilon \cdot \beta_{i,j}$. (One can easily check that the stationary measure for this

Markov chain is independent of ϵ for all small $\epsilon > 0$, and also that our assumptions imply that the transition matrix is irreducible and hence has a unique stationary measure.)

If $m > 2$, the analogue of Theorem 3.2 says only that the eigenfunctions for T_ϵ whose eigenvalues approach 1, but are distinct from 1, limit on the space of linear combinations of invariant densities for T_0 with integral 0.

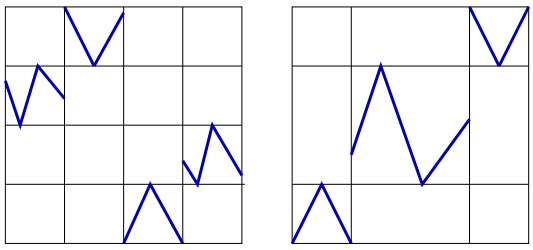

Figure 3.5: Initial maps T_0 which give rise to metastable systems for which our results can be generalized. The initially invariant sets are $I_1 = [0, 1/4] \cup [1/2, 3/4]$ and $I_2 = [1/4, 1/2] \cup [3/4, 1]$ (left) and $I_1 = [0, 1/4]$, $I_2 = [1/4, 3/4]$ and $I_3 = [3/4, 1]$ (right).

Boundary condition.

The restriction that the boundary point does not move when T_0 is perturbed is inessential; when it is relaxed, it simply means that the metastable states for T_ϵ are slight perturbations of the initial invariant sets. In this case, a smooth change of coordinates restores the hypothesis (P2). For example, when $b \notin \mathcal{C}_0$ assumption (P2a) is actually superfluous, although the definitions in the statement of Theorem 3.1 must be modified slightly. As remarked earlier, necessarily $T_0(b) = b$.

Furthermore, the graph of T_0 intersects the diagonal transversely at this point. Thus for all small $\epsilon > 0$, there is a unique point b_ϵ near b satisfying $T_\epsilon(b_\epsilon) = b_\epsilon$. Then the quasi-invariant sets for T_ϵ are $I_{l,\epsilon} := [0, b_\epsilon]$ and $I_{r,\epsilon} := [b_\epsilon, 1]$, and the corresponding holes are defined by $H_{l,\epsilon} := I_{l,\epsilon} \cap T_\epsilon^{-1}(I_{r,\epsilon})$ and $H_{r,\epsilon} := I_{r,\epsilon} \cap T_\epsilon^{-1}(I_{l,\epsilon})$. Aside from these minor modifications, the statements and proofs of our main results remain the same.

When $b \in \mathcal{C}_0$, (P2b) can be relaxed by no longer requiring that $b \in \mathcal{C}_\epsilon$ for all ϵ. In this case, when $\epsilon > 0$, $\mathcal{C}_\epsilon$ contains a point b_ϵ that converges to b as $\epsilon \to 0$, and the quasi-invariant sets and holes must be redefined as above. However, it is still essential to assume that no holes are created near the boundary, which we enforce with the assumption that $T_0(b_-) < b < T_0(b_+)$. For example, if $T_\epsilon(x) = [(3x \mod 1/2)+3\epsilon]\cdot 1_{x<1/2}+[(-3x \mod 1/2)+1/2-\epsilon]\cdot 1_{x>1/2}$, then all of our assumptions aside from (P2) hold, with $b = 1/2$, $\mu_* = \text{Leb}|_{I_*}$, and $l.h.r. = 1/3$. However, as $\epsilon \to 0$, $\phi_\epsilon \xrightarrow{L^1} \phi_r$. The difficulty is that orbits ejected from I_r by T_ϵ immediately return to I_r.

Multiple limiting densities.

When the limit $l.h.r.$ in §3.2.2 does not exist, we let

$$\underline{l.h.r.} = \liminf_{\epsilon\to 0} \frac{\mu_r(H_{r,\epsilon})}{\mu_l(H_{l,\epsilon})}, \quad \overline{l.h.r.} = \limsup_{\epsilon\to 0} \frac{\mu_r(H_{r,\epsilon})}{\mu_l(H_{l,\epsilon})}.$$

Since the function $\frac{\mu_r(H_{r,\epsilon})}{\mu_l(H_{l,\epsilon})}$ is continuous in $\epsilon > 0$, our arguments show that the set of limit points for ϕ_ϵ as $\epsilon \to 0$ is precisely

$$\left\{ \tilde{\alpha}\phi_l + (1-\tilde{\alpha})\phi_r : \frac{\tilde{\alpha}}{1-\tilde{\alpha}} \in [\underline{l.h.r.}, \overline{l.h.r.}] \right\}.$$

3.3 Proofs of the main theorems

In this section, we state the main properties of the invariant densities of the initial system and its perturbations. Then, we present the proofs of Theorems 3.1 and 3.2. For notational convenience, we will assume that there are only two infinitesimal holes, $h_l \in I_l$ and $h_r \in I_r$; the proof without this restriction is essentially unchanged.

3.3.1 Properties of the invariant densities

Here we record some of the relevant characteristics of the density functions $\phi_\epsilon, \phi_l, \phi_r$. First, if $f \in BV$, we can – and will – choose a representative of f with only regular discontinuities, i.e. for each x, $f(x) = (\lim_{y\to x^-} f(y) + \lim_{y\to x^+} f(y))/2$. Then, following [Bal07], we can uniquely decompose $f = f^{reg} + f^{sal}$ into the sum of a regular and a singular (or saltus) part. Here f^{reg} is continuous with $\mathrm{var}(f^{reg}) \leq \mathrm{var}(f)$, and f^{sal} is the sum of at most countably many step functions. We write $f^{sal} = \sum_{u\in\mathcal{S}} s_u H_u$, where $\mathcal{S}$ is the discontinuity set of f, s_u is the *jump* of f at u, and $H_u(x) = -1$ if $x < u$, $-\frac{1}{2}$ if $x = u$ and 0 if $x > u$. This representation imposes the boundary condition $f^{sal}(1) = 0$. Furthermore, $\mathrm{var}(f^{sal}) = \sum_{u\in\mathcal{S}} |s_u| \leq \mathrm{var}(f)$.

Proposition 3.3.1 (Key facts about the invariant densities). There exists $\epsilon_0 > 0$ such that:

(i) *Uniform bound on the variations of the invariant densities.*

$$\sup_{0<\epsilon<\epsilon_0} \mathrm{var}(\phi_\epsilon) < +\infty.$$

Also, $\text{var}(\phi_l), \text{var}(\phi_r) < +\infty$.

(ii) *Uniform bound on the Lipschitz constant of the regular parts.*

For $0 < \epsilon < \epsilon_0$, each of the ϕ_ϵ^{reg} is Lipschitz continuous with constant $\text{Lip}(\phi_\epsilon^{reg})$, and

$$\sup_{0<\epsilon<\epsilon_0} \text{Lip}(\phi_\epsilon^{reg}) < +\infty.$$

Also, ϕ_l^{reg} and ϕ_r^{reg} are Lipschitz.

(iii) *Approximate continuity near the infinitesimal holes.*

For $* \in \{l, r\}$, for each $\eta > 0$, there exists $\delta > 0$ such that for all $0 < \epsilon < \epsilon_0$,

$$\text{var}_{[h_*-\delta, h_*+\delta]}(\phi_\epsilon^{sal}) := \text{the variation of } \phi_\epsilon^{sal} \text{ over } [h_* - \delta, h_* + \delta] < \eta.$$

Also, ϕ_* is continuous at h_*.

The proof of Proposition 3.3.1 is technical, and so we defer it until §3.4.2.

3.3.2 Proofs

We recall that for any $C_1, C_2 > 0$, $\{f \in BV : |f|_{L^1} \leq C_1, \text{var}(f) \leq C_2\}$ is pre-compact in L^1. This fact will be used repeatedly in what follows.

Proof of Theorem 3.1

Using (i) of Proposition 3.3.1, we are able to choose a sequence of values ϵ' converging to 0 such that $\phi_{\epsilon'}$ converges in L^1 to some function, which we denote by ϕ_0. Using the fact that ϕ_ϵ is a fixed point of the Perron-Frobenius operator $\mathcal{L}_\epsilon$ associated to T_ϵ (see §3.4.1 for the definition), one can verify that ϕ_0 is an invariant

density for T_0, and so there exists α such that $\phi_0 = \alpha\phi_l + (1-\alpha)\phi_r$. We will verify that necessarily $\alpha/(1-\alpha) = l.h.r.$. From this it follows that there is exactly one limit point of ϕ_ϵ as $\epsilon \to 0$, and Theorem 3.1 follows.

Now for $\epsilon' > 0$, $\mu_{\epsilon'}(H_{l,\epsilon'}) = \mu_{\epsilon'}(H_{r,\epsilon'})$, because $I_l = H_{l,\epsilon'} \cup (I_l \setminus H_{l,\epsilon'})$ and $T_{\epsilon'}^{-1}(I_l) = H_{r,\epsilon'} \cup (I_l \setminus H_{l,\epsilon'})$ are disjoint unions (modulo sets of zero Lebesgue measure), and $\phi_{\epsilon'} = d\mu_{\epsilon'}/dx$ is an invariant density for $T_{\epsilon'}$. We will show that as $\epsilon' \to 0$,

$$\mu_{\epsilon'}(H_{l,\epsilon'}) = \alpha\mu_l(H_{l,\epsilon'}) + o(1)\cdot\mu_l(H_{l,\epsilon'}), \tag{3.1}$$

$$\mu_{\epsilon'}(H_{r,\epsilon'}) = (1-\alpha)\mu_r(H_{r,\epsilon'}) + o(1)\cdot\mu_r(H_{r,\epsilon'}), \tag{3.2}$$

from which the equation $\alpha/(1-\alpha) = l.h.r.$ and hence Theorem 3.1 follows immediately.

We prove only Equation (3.1), since the proof of Equation (3.2) is analogous. Write

$$\begin{aligned}\mu_{\epsilon'}(H_{l,\epsilon'}) &= \int_{H_{l,\epsilon'}} \phi_{\epsilon'}\,dx = \alpha\int_{H_{l,\epsilon'}} \phi_l\,dx + \int_{H_{l,\epsilon'}} (\phi_{\epsilon'} - \alpha\phi_l)\,dx \\ &= \alpha\mu_l(H_{l,\epsilon'}) + O\left(\sup_{x\in H_{l,\epsilon'}} |\phi_{\epsilon'}(x) - \alpha\phi_l(x)|\right)\cdot \mathrm{Leb}(H_{l,\epsilon'}).\end{aligned}$$

But as $\epsilon' \to 0$, $H_{l,\epsilon'} \to h_l$ in the Hausdorff metric, and then $\mu_l(H_{l,\epsilon'})/\mathrm{Leb}(H_{l,\epsilon'}) \to \phi_l(h_l) > 0$, because ϕ_l is continuous at h_l. Thus our proof is completed by the following:

Lemma 3.3.2. As $\epsilon' \to 0$,

$$\sup_{x\in H_{l,\epsilon'}} |\phi_{\epsilon'}(x) - \alpha\phi_l(x)| \to 0.$$

Although this uniform convergence might at first seem surprising, Proposition 3.3.1 (ii) and (iii) essentially say that near h_l, $\{\phi_{\epsilon'}\}$ behaves like a family of

equicontinuous functions.

Proof. We proceed by contradiction. Suppose that there exists $C > 0$ and a subsequence $\epsilon'' \to 0$ of the ϵ' values such as that for each ϵ'', there is a point $x_{\epsilon''} \in H_{l,\epsilon''}$ with $|\phi_{\epsilon''}(x_{\epsilon''}) - \alpha\phi_l(x_{\epsilon''})| > C$. Necessarily, $x_{\epsilon''} \to h_l$ as $\epsilon'' \to 0$.

We restrict all functions of interest to the left subinterval I_l. Set $\gamma_{\epsilon''} := \phi^{reg}_{\epsilon''} - \alpha\phi^{reg}_l$ and $\omega_{\epsilon''} := \phi^{sal}_{\epsilon''} - \alpha\phi^{sal}_l$, so that $\phi_{\epsilon''} - \alpha\phi_l = \gamma_{\epsilon''} + \omega_{\epsilon''}$. Using (ii) of Proposition 3.3.1, let L be such that for all sufficiently small ϵ'', $\mathrm{Lip}(\gamma_{\epsilon''}) < L$. Next, we use (iii) with $\eta = C/5$ and make sure to choose the corresponding $\delta < C/(5L)$ small enough so that $\mathrm{var}_{[h_l-\delta,h_l+\delta]}(\alpha\phi^{sal}_l) < C/5$ as well. Thus $\mathrm{var}_{[h_l-\delta,h_l+\delta]}(\omega_{\epsilon''}) < 2C/5$. Then if $x \in [h_l - \delta, h_l + \delta]$, and ϵ'' is sufficiently small, $x_{\epsilon''} \in [h_l - \delta, h_l + \delta]$ and

$$\begin{aligned}
&|\gamma_{\epsilon''}(x) + \omega_{\epsilon''}(x)| \\
&\geq |\gamma_{\epsilon''}(x_{\epsilon''}) + \omega_{\epsilon''}(x_{\epsilon''})| - |\gamma_{\epsilon''}(x) + \omega_{\epsilon''}(x) - \gamma_{\epsilon''}(x_{\epsilon''}) - \omega_{\epsilon''}(x_{\epsilon''})| \\
&\geq C - [L \cdot 2\delta + 2C/5] \geq C/5.
\end{aligned}$$

But this contradicts that $\gamma_{\epsilon''} + \omega_{\epsilon''} = \phi_{\epsilon''} - \alpha\phi_l \xrightarrow{L^1} 0$. □

Proof of Theorem 3.2

First, we observe that the results of [KL99] guarantee that for small $\epsilon > 0$, $\rho_\epsilon < 1$ is a simple eigenvalue of multiplicity 1. Hence there are exactly two real-valued eigenfunctions, $\pm\psi_\epsilon$, satisfying $\mathcal{L}_\epsilon\psi_\epsilon = \rho_\epsilon\psi_\epsilon$ and $|\psi_\epsilon|_{L^1} = 1$. But for such functions, $\int \psi_\epsilon\, dx = \int \mathcal{L}_\epsilon\psi_\epsilon\, dx = \rho_\epsilon \int \psi_\epsilon\, dx$, so $\int \psi_\epsilon\, dx = 0$. We have the following uniform bound on their variations, whose proof we defer until §3.4.2.

Lemma 3.3.3 (Uniform bound on the variations of the ψ_ϵ). There exists $\epsilon_1 > 0$ such that $\sup_{0<\epsilon<\epsilon_1} \mathrm{var}(\psi_\epsilon) < +\infty$.

Let ψ_0 be any limit point in L^1 of ψ_ϵ as $\epsilon \to 0$. Then, since $\rho_\epsilon \to 1$, it follows that ψ_0 is invariant under $\mathcal{L}_0$, and is thus a linear combination of ϕ_l and ϕ_r. Since $|\psi_0|_{L^1} = 1$ and $\int \psi_0 \, dx = 0$, necessarily $\psi_0 = \pm\frac{1}{2}\phi_l \mp \frac{1}{2}\phi_r$. Hence we can uniquely specify ψ_ϵ by the condition $\int_{I_l} \psi_\epsilon \, dx > 0$, and Theorem 3.2 follows.

3.4 Proofs of the properties of the densities

In order to prepare for the proofs of Proposition 3.3.1 and Lemma 3.3.3, it will be convenient to first show how to derive such properties for an invariant density of a single, fixed piecewise expanding map. We do this in §3.4.1. Then, in §3.4.2, we prove Proposition 3.3.1 and Lemma 3.3.3 by showing how such estimates can be made uniformly for the family of maps T_ϵ, $\epsilon \geq 0$.

Before beginning, we remark that if $f \in BV$, then for each x, $|f|_\infty \leq |f(x)| + \mathrm{var}(|f|) \leq |f(x)| + \mathrm{var}(f)$. Integrating, we find that $|f|_\infty \leq |f|_{L^1} + \mathrm{var}(f)$. We will use this fact repeatedly below.

3.4.1 Properties of an invariant density for a single piecewise expanding map

Let $T : I \circlearrowleft$ be a piecewise C^2 uniformly expanding map, with $\mathcal{C} = \{0 = c_0 < c_1 < \cdots < c_d = 1\}$ as a critical set. Let $\mathcal{L}$ be the associated Perron-Frobenius operator, i.e., the transfer operator acting on densities. We begin by briefly re-

viewing a method for finding an invariant density of T. Such a method was introduced in [LY73]; see Chapter 3 in [Bal00] for a more modern exposition. Let $\lambda_T = \inf_{x \in I \setminus \mathcal{C}} |T'(x)| > 1$ be the minimum expansion and $D_T = \sup_{x \notin \mathcal{C}} |T''(x)| / |T'(x)|$ be the distortion of T. Then if $f \in BV$, $x \notin T\mathcal{C}$,

$$\mathcal{L}f(x) = \sum_{i=1}^{d} f(\xi_i(x)) \, |\xi_i'(x)| \, 1_{J_i}(x), \tag{3.3}$$

where $J_i = T|_{[c_{i-1},c_i]}([c_{i-1}, c_i])$ and $\xi_i = (T|_{[c_{i-1},c_i]})^{-1} : J_i \to [c_{i-1}, c_i]$. One can show that there exists constants $\beta \in (0,1)$ and C_{LY} such that for each $n \geq 1$ and $f \in BV$, the following Lasota-Yorke inequality holds:

$$\mathrm{var}(\mathcal{L}^n f) \leq C_{\mathrm{LY}} \beta^n \mathrm{var}(f) + C_{\mathrm{LY}} \, |f|_{L^1} . \tag{3.4}$$

In fact, β can be chosen as any number greater than λ_T^{-1}, although we will not use this fact. Set $F_n = \frac{1}{n} \sum_{k=0}^{n-1} \mathcal{L}^k 1$. Then $F_n \xrightarrow{L^1} \phi$, where $\phi \in BV$ is the density of an ACIM for T. Using Helly's Theorem, one has that $\mathrm{var}(\phi) \leq C_{\mathrm{LY}}$.

We wish to characterize the properties of the regular and singular terms in the decomposition $\phi = \phi^{reg} + \phi^{sal}$. First, let us define a hierarchy on the set of points in the postcritical orbits $\mathcal{S} = \cup_{k \geq 1} T^k \mathcal{C}$ by $\#(u) := \inf\{k \geq 1 : u \in T^k \mathcal{C}\}$. The following characterization is motivated by the discussion of the invariant densities for unimodal expanding maps found in [Bal07] and [BS08]. In particular, in [BS08, §3.3] a norm is introduced on the sequence of jumps of ϕ along the postcritical orbit with weights that grow exponentially in $\#(u)$.

Lemma 3.4.1. Given the hypotheses above,

(a) ϕ^{reg} is Lipschitz continuous. Furthermore, there exists a constant $C_{\mathrm{dis}} =$

$C_{\text{dis}}(\lambda_T, D_T)$ such that $\text{Lip}(\phi^{reg}) \le C_{\text{dis}}(1 + C_{\text{LY}})$. $C_{\text{dis}}(\lambda_T, D_T)$ can be defined so that it depends continuously on $\lambda_T > 1$, $D_T \ge 0$.

(b) The discontinuity set of ϕ is a subset of $\mathcal{S} = \cup_{k\ge 1} T^k \mathcal{C}$. If we write $\phi^{sal} = \sum_{u\in\mathcal{S}} s_u H_u$, then for each $m \ge 0$, $\sum_{\{u\in\mathcal{S}:\#(u)>m\}} |s_u| \le \lambda_T^{-m} C_{\text{LY}}$.

Proof. We begin by noting from Equation (3.3) that F_n is smooth except possibly at points in $\cup_1^{n-1} T^k \mathcal{C}$. Write $F_n^{sal} = \sum_{u\in\mathcal{S}} s_{u,n} H_u$. Then we can show that $|s_{u,n}|$ decays uniformly exponentially fast in $\#(u)$, i.e.

Sublemma. For each $m, n \ge 0$, $\sum_{\{u\in\mathcal{S}:\#(u)>m\}} |s_{u,n}| \le \lambda_T^{-m} C_{\text{LY}}$.

Proof of the Sublemma. If $m \ge n$, $\sum_{\#(u)>m} |s_{u,n}| = 0$, and $\sum_{\#(u)>0} |s_{u,n}| = \text{var}(F_n^{sal}) \le \text{var}(F_n) \le C_{\text{LY}}$. Since $F_n = \frac{n-1}{n}\mathcal{L}F_{n-1} + \frac{1}{n}$, if $\#(u) > 1$ we see from Equation (3.3) with $f = F_{n-1}$ that $|s_{u,n}| \le \frac{n-1}{n}\lambda_T^{-1} \sum_{\{v\in\mathcal{S}:Tv=u\}} |s_{v,n-1}|$. Thus if $0 < m < n$,

$$\begin{aligned}\sum_{\#(u)>m} |s_{u,n}| &\le \sum_{\#(u)>m} \frac{n-1}{n}\lambda_T^{-1} \sum_{\{v\in\mathcal{S}:Tv=u\}} |s_{v,n-1}| \\ &\le \frac{n-1}{n}\lambda_T^{-1} \sum_{\#(u)>m-1} |s_{u,n-1}| \le \cdots \\ &\le \frac{n-m}{n}\lambda_T^{-m} \sum_{\#(u)>0} |s_{u,n-m}| \le \lambda_T^{-m} C_{\text{LY}}.\end{aligned}$$

In the inequalities above, we use the fact that if $\#(u) > 1$, then $T^{-1}(u)$ does not contain any critical points. □

Using a diagonalization argument, we may find a subsequence n_j such that for each u, s_{u,n_j} converges as $n_j \to \infty$ to some number, which we write as $\hat{s}_u$. In particular, for each m, $\sum_{\#(u)>m} |\hat{s}_u| \le \lambda_T^{-m} C_{\text{LY}}$, and $F_{n_j}^{sal} \xrightarrow{L^1} F^{sal}$, where we define $F^{sal} = \sum_{u\in\mathcal{S}} \hat{s}_u H_u$. Furthermore, a standard distortion estimate (see for example

the proof of Proposition 3.3 in [Bal07]) shows that there exists a constant C_{dis} such that for each $n \geq 0$, $\text{Lip}((\mathcal{L}^n 1)^{reg}) \leq C_{\text{dis}} \left|\mathcal{L}^n 1\right|_\infty \leq C_{\text{dis}}(1 + \text{var}(\mathcal{L}^n 1))$. Here, C_{dis} depends only on the minimum expansion and on the distortion of T. In particular, $\sup_{n\geq 1} \text{Lip}(F_n^{reg}) \leq C_{\text{dis}}(1 + C_{\text{LY}})$. By the Arzelà-Ascoli Theorem, we may find a continuous function F^{reg} such that some subsequence of $\{F_{n_j}^{reg}\}$ converges in L^∞ to F^{reg}.

By the uniqueness of the decomposition $\phi = \phi^{reg} + \phi^{sal}$, we conclude that $\phi^{reg} = F^{reg}$ and $\phi^{sal} = F^{sal}$. Lemma 3.4.1 follows.

□

3.4.2 Proofs of Proposition 3.3.1 and Lemma 3.3.3

We prove only the claims about ϕ_ϵ for $\epsilon > 0$, and leave the claims about ϕ_l, ϕ_r to the reader.

Let $\mathcal{L}_\epsilon$ be the Perron-Frobenius operator (3.3) associated to T_ϵ. The first key step is to prove that the $\mathcal{L}_\epsilon$ with ϵ sufficiently small satisfy Lasota-Yorke inequalities with uniform constants. Let λ_ϵ and D_ϵ be the minimum expansion and distortion of T_ϵ, respectively. Then as $\epsilon \to 0$, $\lambda_\epsilon \to \lambda_0$ and $D_\epsilon \to D_0$. Furthermore, T_ϵ is a piecewise C^2 uniformly expanding map that is a small C^2 perturbation of T_0, and the two critical sets $\mathcal{C}_\epsilon$, $\mathcal{C}_0$ are $\epsilon-$close together. This is not sufficient to guarantee uniform Lasota-Yorke inequalities, see for example [Kel82, §6] or [Bla92]. However, such uniform inequalities do follow with the additional assumption (I4), which guarantees that either (a) we have $\lambda_0 > 2$ or (b) T_0 has no periodic critical points,

except possibly the points in ∂I as fixed points. We assume the former case in our presentation here, and comment on the latter case at the end of this section.

Fix $\lambda \in (2, \lambda_0)$. The original proof from [LY73] shows that if $f \in BV$ is real-valued,

$$\mathrm{var}(\mathcal{L}_\epsilon f) \leq (2\lambda_\epsilon^{-1})\mathrm{var}(f) + C_\epsilon \, |f|_{L^1} \, ,$$

where

$$C_\epsilon = D_\epsilon/\lambda_\epsilon + 2 \max_i |c_{i+1,\epsilon} - c_{i,\epsilon}|^{-1} \, . \tag{3.5}$$

(Compare also [Liv95][§2].) Iterating, we find that for sufficiently small ϵ, for all such f and $n \geq 1$,

$$\mathrm{var}(\mathcal{L}_\epsilon^n f) \leq \beta^n \mathrm{var}(f) + C_{\mathrm{LY}} \, |f|_{L^1} \, , \tag{3.6}$$

with $\beta = 2\lambda^{-1}$ and $C_{\mathrm{LY}} = 2C_0/(1 - 2\lambda^{-1})$. Similar estimates can be made for complex-valued f by applying (3.6) to the real and imaginary parts separately. Since each T_ϵ has a unique ACIM, we know from our discussion in §3.4.1 that for sufficiently small $\epsilon > 0$, $\frac{1}{n}\sum_{k=0}^{n-1} \mathcal{L}_\epsilon^k 1 \xrightarrow{L^1} \phi_\epsilon$ as $n \to \infty$. It follows from Lemma 3.4.1 that $\mathrm{var}(\phi_\epsilon)$ and $\mathrm{Lip}(\phi_\epsilon^{reg})$ are uniformly bounded.

Next, we prove (iii). Given $\eta > 0$, choose n large enough that $\lambda^{-n} C_{\mathrm{LY}} < \eta$. Using (I2), we can choose $\delta > 0$ so small that for $0 < k \leq n$, $(T_0^k \mathcal{C}_0) \cap [h_* - 2\delta, h_* + 2\delta] = \emptyset$. It follows that for ϵ sufficiently small, $(T_\epsilon^k \mathcal{C}_\epsilon) \cap [h_* - \delta, h_* + \delta] = \emptyset$ as well. Using part (b) of Lemma 3.4.1 with $m = n$, we then see that $\mathrm{var}_{[h_*-\delta, h_*+\delta]}(\phi_\epsilon^{sal}) < \eta$.

Finally, to prove Lemma 3.3.3, we use Equation (3.6) with $f = \psi_\epsilon$, n chosen so large that $\beta^n < 1/2$, and ϵ chosen so small that $\rho_\epsilon^n > 3/4$. It follows that $\mathrm{var}(\psi_\epsilon) \leq C_{\mathrm{LY}}/(\rho_\epsilon^n - \beta^n) \leq 4C_{\mathrm{LY}}$.

Modifications when the minimum expansion is not bigger than two

If, in assumption (I4), the minimum expansion of T_0 is $\lambda_0 \leq 2$, one derives Lasota-Yorke estimates for $\mathcal{L}_0$ by first fixing N large enough so that $\lambda_0^N > 2$. Then the arguments from [LY73] used above will yield a Lasota-Yorke estimate for $\mathcal{L}_0^N$, and this can be interpolated to give similar estimates for $\mathcal{L}_0$. One can try to obtain uniform estimates for $\mathcal{L}_\epsilon$, but the arguments used above will only work if the critical points for T_ϵ^N are in a one-to-one correspondence with and very close to those of T_0^N, compare Equation (3.5), as would be the case if $\mathcal{C}_0 \cap (\cup_{k=1}^{N-1} T^k \mathcal{C}_0) = \emptyset$.

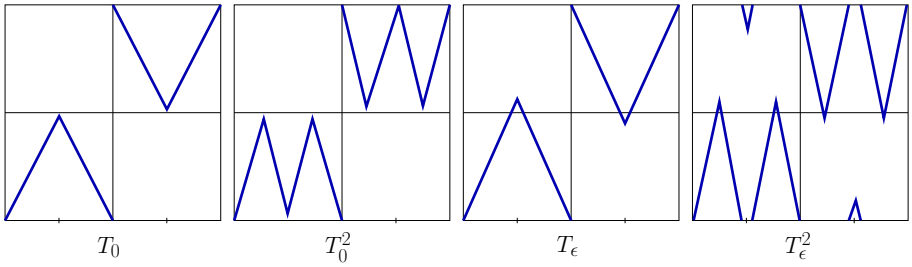

Figure 3.6: Creation of small intervals of differentiability.

Unfortunately, this will never be the case in our setting, at least when $b \in \mathcal{C}_0$. This is because the infinitesimal holes in H_0 are necessarily critical points, and they are mapped to b by T_0. Because at least some of the infinitesimal holes must be mapped across the boundary point when $\epsilon > 0$, this means that necessarily T_ϵ^2 will have more critical points than T_0^2, and these additional critical points will create very short intervals on which T_ϵ^2 is smooth; see Figure 3.6. However, this problem can be dealt with using assumption (I4b). Specifically, in [BY93] it is shown that because of the restriction on the periodic critical points, the growth in the number

of the very short intervals on which T_ϵ^n is smooth as n increases can be controlled, and that uniform Lasota-Yorke estimates can still be made. Precisely, there exists $\epsilon_0 > 0$ and constants $\beta \in (0,1)$, C_{LY} such that for each $\epsilon \in [0, \epsilon_0]$, $n \geq 1$ and $f \in BV$,

$$\mathrm{var}(\mathcal{L}_\epsilon^n f) \leq C_{\mathrm{LY}} \beta^n \mathrm{var}(f) + C_{\mathrm{LY}} |f|_{L^1} . \tag{3.7}$$

The proof of this is essentially identical to the proof of Lemma 3.2 in [Bal00] (see also her Remark 3.4, and compare the proof of Lemma 8 in [BY93]), and so we omit it. The rest of the proofs of Proposition 3.3.1 and Lemma 3.3.3 proceed as above.

Chapter 4

A data assimilation method for hyperbolic systems

4.1 Introduction

In data assimilation, algorithms to best combine data collected from measurements with forecasts from a mathematical model are sought. See [Kal02] for an overview of data assimilation methods. In ensemble data assimilation, we work with a set of model state vectors, called an *ensemble*, that intends to describe and keep track of position and uncertainty of the system state. See [Eve09] for an overview of ensemble data assimilation.

Each data assimilation cycle consists of two steps: forecast and analysis steps. In the ensemble approach, the forecast step takes as initial conditions the *analysis* ensemble from the previous cycle and evolves each ensemble member separately according to an appropriate model, generating the *background* ensemble. Then, the information collected from measurements is used to produce the new analysis ensemble, by adjusting the background ensemble toward the data observed. Since the model state may not be measured directly, in general the analysis step uses an *observation function* (also called a *forward operator*) that quantifies what the measurements should be for a given model state. At this step, data is filtered according to the algorithm, and adjustments are made along a space determined by the ensemble members.

An example where data assimilation is used heavily is in forecasting the weather. This is the motivation behind the present work. Weather models are very high dimensional, as the state of the system comprises information about the Earth's entire atmosphere. At the resolutions currently used, there are millions of variables involved. The main goal is to predict the state of the system in the future; that is, weather forecasting. Data assimilation is required because the current state of the system is not well determined by current observations. In this context, observations are measurements meteorological variables, such as temperature, at various locations. A successful data assimilation procedure combines information collected from the observations with the forecast generated by the weather model, and produce, at each step, a good approximation to the corresponding state of the system.

In more general terms, a problem of interest in data assimilation is to identify a trajectory of a dynamical system that produces a given sequence of observations (time series), whether or not the ultimate goal is predicting the future behavior of the system.

In this paper, we investigate dynamical properties of a data assimilation algorithm, an ensemble Kalman filter (EKF) studied in [HKS07], assuming the underlying system is uniformly hyperbolic. The Kalman filter was introduced in [Kal60]. It is optimal, in a least square sense, for the case of linear model and observations with white noise. Extensions to the non-linear setting (Extended KF) have been developed, see for example [RGYU99]. They involve linearizations and model-size matrix inversions, thus making computations costly for high-dimensional systems.

Ensemble Kalman filters were introduced in [Eve94], and further developed and tested in [BvLE98, HM98]. They keep track of a set (ensemble) of trajectories, and are suitable for parallel computations. More recently, methods with deterministic choice of ensemble elements have been developed [And01, BEM01, OHS+04, TAB+03, WBJ04, WH02]. The algorithm studied here belongs to this class. While we consider a particular algorithm, we believe our methods of proof can be adapted to other deterministic EKF.

In order to track a system $f : M \circlearrowleft$, usually referred to as the *truth* and known only to within some accuracy, it is necessary to be able to adjust forecasts along all unstable directions. The number of these may be much smaller than the total dimension of the system. In this case, we can hope to keep track of the truth by keeping track of the evolution of an ensemble of trajectories surrounding an approximation to the truth, with the number of elements in the ensemble related to the number of unstable directions. Hyperbolic systems possess a well defined number of unstable directions, independent of the trajectory. For this reason, they provide a tractable setting to investigate the properties of EKF. Because the differences between ensemble members determine directions in which EKF can make adjustments, the number of ensemble members must exceed the dimension of the unstable space in order to be able to correct errors in all unstable directions.

A fundamental property of hyperbolic systems is shadowing. This property ensures that for any $\delta > 0$ there exists some $\epsilon > 0$ such that every $\epsilon-$pseudo-orbit of the system is $\delta-$shadowed by a real orbit. An $\epsilon-$pseudo-orbit of f is a sequence $\{x_n\}_{a<n<b} \subset M$ for which $\|f(x_n) - x_{n+1}\| < \epsilon$ for all $a < n < b$, and it is said to be

$\delta-$shadowed by the orbit of x if $\|x_n - f^n(x)\| < \delta$ for all $a < n < b$. See [KH95, §18] for a precise statement of the shadowing lemma.

The main result of the paper, presented in Proposition 4.3.1 and generalized in Proposition 4.4.4, ensures that for hyperbolic attractors with $k^u < k$ dimensional unstable spaces, the shadowing property holds for a non-empty open set of initial k-member EKF, under Takens' genericity conditions [Tak81] for the observation function. This property guarantees that the data assimilation procedure is reliable, in the sense that when appropriately initialized, its trajectory provides an approximation for the true trajectory to within a small error for all future time. In other words, the data assimilation system, driven by the real system, synchronizes with it. Consequences for the approximation of positive Lyapunov exponents of the system are also presented.

The defining property of unstable spaces, and more precisely the existence of invariant unstable cones in some dynamical systems, suggests a convenient way of identifying meaningful ensemble members, by thinking of the differences between ensemble members as vectors in the tangent space. Namely, if we intend to track the dynamics near a particular trajectory, we may keep track of the evolution of tangent vectors under the (tangent) dynamics. A consequence of the theorem of Oseledec [Ose68] is that for almost every initial trajectory (with respect to an invariant measure for the system) almost every vector approaches the most unstable direction for that trajectory. If we are careful to orthogonalize and normalize the vectors at each step we can identify the other unstable spaces in the Oseledec filtration as well, through the standard numerical procedure for estimating Lyapunov

exponents; see the Appendix of [SN79]. In [GPT+07] an alternative algorithm to identify spaces of the Oseledec splitting is introduced.

The task we just described, including the computation of the tangent map, may be costly, computationally or otherwise. In fact, considerable human time is devoted to linearizing weather models, see p. 215 and Appendix B of [Kal02]. Moreover, dealing with derivatives significantly increases, at the least, storage requirements. As an alternative, we can evolve ensemble vectors according to f instead of Df. This is analogous to using the secant method instead of Newton's method as a root-finding algorithm in calculus or numerical analysis. If the size of the ensemble vectors is small, of order ϵ, the distance between the image of a point and its linearization is of order ϵ^2.[1] We would like to show that under some circumstances, this procedure indeed produces ensemble vectors that lie inside an unstable cone. When this is the case, if the cones are strictly invariant, once inside the cones, the algorithm would keep successive iterates of the ensemble inside unstable cones. On the one hand, this would allow to adjust errors accumulated in unstable directions. On the other hand, it would permit to iterate the procedure.

It would be of great interest to find extensions of our results to more general underlying dynamical systems. For example, non-uniformly hyperbolic systems are believed to provide adequate models for several phenomena in the natural sciences. They also share some properties with the systems treated here, such as existence of stable and unstable manifolds for Lyapunov regular points, and some of them also

[1] In general, it would be necessary to use the exponential map to identify tangent vectors with points in the space.

have complete strict (or eventually strict) invariant families of cones as described in [BP06]. However, in general, stable and unstable spaces do not depend continuously on the base point. Hence our proofs do not extend directly to that setting, and other techniques would be necessary to study the problem in such generality.

The structure of the paper is as follows. In §4.2, we present the main result, after introducing the setting and discussing the initialization of the ensemble. In §4.3, properties of the ensemble Kalman filter are established for the case of one-dimensionally unstable hyperbolic systems. These properties are generalized to the case of higher dimensionally unstable hyperbolic systems in §4.4. The main reason to separate the two cases is to present the control of nonlinear terms in §4.3, and leave the main complications of the extension to higher dimensionally unstable cases, which lie at the linear level, to §4.4.

4.2 Statement of results

4.2.1 Setting

We start by describing our hypotheses for the forecast model and the observation function.

4.2.1.1 Model

Throughout the paper, let $f : M \circlearrowleft$ be a C^3 diffeomorphism of a Riemannian manifold[2] of dimension N, with a uniformly hyperbolic attractor of $\mathcal{A} \subset M$. That is, a compact set invariant under f for which there is an open set $U \subset M$ such that $\mathcal{A} \subset int(U)$ and $\cap_{n\geq 0} f^n(U) = \mathcal{A}$. The hyperbolicity condition means that, restricted to $\mathcal{A}$, there is an f invariant splitting of the tangent spaces T_xM into unstable (expanding) and stable (contracting) spaces, $T_xM = E^u_x \oplus E^s_x$, and constants $\lambda > 1 > \mu$ such that

$$\|Dfv\| > \lambda\|v\| \quad \forall v \in E^u \text{ and } \|Dfv\| < \mu\|v\| \quad \forall v \in E^s.$$

A reference for hyperbolic systems is [KH95].

Remark 4.2.1. This is not the usual definition of hyperbolicity, in the sense that we are already working with a metric adapted to f. This assumption simplifies some calculations, but does not restrict the scope of the paper since adapted metrics always exist. Moreover, even if a metric (e.g. Euclidean) is not adapted to a hyperbolic system f, it will be adapted to a suitable power.

For the reminder of the paper, we also assume that for each f periodic point x of period $k \leq 2N + 1$ the eigenvalues of Df^k_x are distinct. This open and dense property in $Diff(M)$ is needed for Takens' embedding theorem to apply; see Theorem 4.2.1.2 for the statement.

[2] To avoid technical difficulties, we assume M is a Euclidean space, a cylinder or a torus so there is no need to make use of the exponential map to identify tangent vectors with points in the space.

4.2.1.2 Observation function

For f fixed, we consider generic C^2 real-valued observation functions $h : M \to \mathbb{R}$ in the sense of the following theorem, which will be repeatedly used in this paper.

Theorem (Takens embedding theorem, [Tak81])**.** Let $f : M \circlearrowleft$ as in §4.2.1.1. Then, for smooth proper[3] functions $h : M \to \mathbb{R}$, it is a generic property that the map

$$x \mapsto \big(h(x), h(f(x)), h(f^2(x)), \ldots, h(f^{2N}(x))\big)$$

is an embedding, i.e. one-to-one proper immersion.

We note that this result has been (or may be) refined in a couple of ways that may be relevant for concrete applications. On the one hand, generalizations of Theorem 4.2.1.2, such as those in [SYC91] may be useful. In short, they allow to reduce the number of measurements from $2N + 1$ to $2 \dim \mathcal{A} + 1$, where $\mathcal{A}$ is the attractor of f under consideration. This would improve the estimates significantly, as errors grow exponentially with the number of steps considered.

On the other hand, there may be multiple observations available at each step, say l scalar measurements. To extend our results to this setting, a multidimensional version of Takens' theorem is needed. Such an extension may be established following Takens' original proof. Therefore our proof could be adapted to, in some appropriate sense, generic $h : M \to \mathbb{R}^l$, and reduce the number of forecast steps considered by a factor of l. It is also possible to reduce to the one-dimensional case

[3] A function is proper if the inverse image of every compact set is compact. This is always the case for smooth functions if the domain is a compact manifold.

by assimilating observations sequentially, as discussed in [May79, §7.4] or [WH02, §3]. We do not present all the details here.

4.2.2 Main result

Definition 1. A DAP with initial ensemble $\mathcal{E}$ is called $c-$*reliable* if the predictions it produces eventually shadow the true trajectory within error c. We say that a family of DAPs depending on a parameter ϵ, and having initial ensembles $\{\mathcal{E}_\epsilon\}_{\epsilon>0}$, is $\mathcal{O}(\epsilon)-$ *reliable* if there is some constant c independent of ϵ such that for all $\epsilon > 0$ sufficiently small, the DAP associated to parameter ϵ with initial ensemble $\mathcal{E}_\epsilon$ is $c\epsilon-$reliable.

Consider $f : M \circlearrowleft$ as in §4.2.1.1 and $h : M \to \mathbb{R}$ as in §4.2.1.2. The main result of this paper concerns the reliability of a family of DAPs associated to f and h, provided they are properly initialized. This family is called the k-member ensemble Kalman filter (kMEKF). The iterative algorithm defining the kMEKF and its correspondence with that of [HKS07] will be discussed in detail in §§4.4.1 and 4.4.2. Here we include the definition for the reader's convenience.

Let $\epsilon > 0$. The kMEKF corresponding to ϵ is constructed using the following iterative procedure. After step $n-1$, we start with an ensemble of k vectors with mean $\overline{x}^a_{n-1}$ and displacement vectors $v^a_{j,n-1}$, $1 \leq j \leq k$. The corresponding background mean at step n is denoted by $\overline{x}^b_n$ and the corresponding displacements by

$v^b_{j,n}$, where

$$\overline{x}^b_n = \frac{1}{k}\sum_{j=1}^{k} f(\overline{x}^a_{n-1} + v^a_{j,n-1}),$$

$$v^b_{j,n} = f(\overline{x}^a_{n-1} + v^a_{j,n-1}) - \overline{x}^b_n.$$

The average value of the measurements of h will be $\overline{h}_n := \frac{1}{k}\sum_{j=1}^{k} h(\overline{x}^b_n + v^b_{j,n})$. The measurement of h from the true trajectory x_n will be denoted by $y_n = h(x_n)$. To define the analysis ensemble at step n, we introduce some further notation. For $1 \leq j \leq k$, let

$$q_{j,n} := \frac{1}{\epsilon}(h(\overline{x}^b_n + v^b_{j,n}) - \overline{h}_n), \quad q^2_n := \frac{1}{(k-1)}\sum_{j=1}^{k} q^2_{j,n},$$

$$\gamma_n := \frac{1}{q^2_n}\left(1 - \frac{1}{\sqrt{1+q^2_n}}\right) \text{ if } q_n \neq 0, \gamma_n = 0 \text{ otherwise, and}$$

$$v_{0,n} := \frac{1}{(k-1)}\sum_{j=0}^{k} q_{j,n} v^b_{j,n}.$$

The analysis ensemble is defined by:

$$\overline{x}^a_n = \overline{x}^b_n + \frac{(y_n - \overline{h}_n)}{1+q^2_n}\frac{v_{0,n}}{\epsilon}, \tag{**}$$

$$v^a_{n,j} = v^b_{n,j} - \gamma_n q_{j,n} v_{0,n}, \text{ for } 1 \leq j \leq k.$$

Let $f : M \circlearrowleft$ be as in §4.2.1.1, and let $\mathcal{A}$ be a k^u dimensionally unstable attractor for f, i.e. $\dim E^u = k^u$. Assume $k > k^u$ and $x_0 \in \mathcal{A}$. Our main result is:

Main Result. For generic observation function h there is a family $\{\mathcal{I}_\epsilon\}_{\epsilon>0}$ of open sets of initial ensembles such that whenever $\mathcal{E}_\epsilon \in \mathcal{I}_\epsilon$, the kMEKF with initial ensembles $\mathcal{E}_\epsilon$ and noiseless observation of h is $\mathcal{O}(\epsilon)-$reliable. The same conclusion holds if the measurement of h has noise, provided its size is bounded by a small multiple

of ϵ, and also when the observations are generated by a pseudo-trajectory, provided its distance to a true trajectory is sufficiently small.

Remark 4.2.2. Sets of initial ensembles for which the kMEKF is $\mathcal{O}(\epsilon)-$reliable are described explicitly in §4.4.3.1. Roughly speaking, they consist of ensembles for which $\|x_0 - \overline{x}_0^a\| \lesssim \sqrt{\epsilon}$ and such that the corresponding perturbations are of adequate spread and lie sufficiently close to the unstable space $E^u_{x_0}$. The last condition is discussed and explained in §4.2.3.

This result is proved using an inductive scheme. In §4.3 it is established for the case $k = 2$ (see Proposition 4.3.1 and Corollary 4.1). The general case is deferred to §4.4 (see Proposition 4.4.4). The proofs follow a similar strategy, but the analysis at the linear level is straightforward in the former. Hence, we concentrate in controlling nonlinear terms in §4.3, and leave the complications at the linear level coming from higher dimensional unstable dynamics for §4.4.

4.2.3 Initialization of the ensemble

In this section we discuss how to identify an *initial* ensemble of trajectories that is appropriate for the EKF to be reliable. Proofs are left for subsequent sections. The most desirable characteristic of an initial ensemble is to well approximate the unstable space of the true trajectory. Even in the case of perfect model, which is the one treated here, this is a non-trivial task, as unstable spaces depend on the infinite future of the system.

An unstable cone at x, K^u_x, is a subset of T_xM of the form $K^u_x = \{(v_u, v_s) \in$

$E_x^u \oplus E_x^s = T_xM | \|v_s\| \leq c\|v_u\|\}$ for some constant $c > 0$. The initialization we propose relies on the forward invariance of unstable cones for uniformly hyperbolic systems. This property ensures the existence of a family of cones $K_x^u \subset T_x(M)$ surrounding the unstable space $E_x^u \subset T_x(M)$ that is invariant under Df. Moreover, in the uniformly hyperbolic setting, the invariance is strict, in the sense that $Df_x K_x^u \subset intK_{f(x)}^u \cup \{0\}$. Thus, an unstable cone at a point x gets mapped inside the interior the corresponding cone at $f(x)$ under the tangent dynamics Df. As we do not make use of the linearization, but of the map itself in the *forecast* step, strict invariance is essential to allow for the small errors associated to this difference to be negligible. This permits to ensure that when the displacements of ensemble vectors from the mean are small and lie inside the unstable cone, so do their corresponding images under f.

The above justifies the existence of an open set of ensembles having the desired property of remaining close to the unstable space under application of the dynamics. However, there is still something to be said about how to identify them. A reasonable approach is as follows. We may start with a cloud of points sufficiently dense in a sphere of small radius around the point x. By forecasting according to f, projecting back to a small sphere around $f(x)$, and repeating this procedure for a few steps, we could identify finite time unstable directions, which necessarily contain unstable cones. In fact, by performing a Gram-Schmidt orthogonalization procedure, we may be able to estimate the dimension of the unstable space. This estimate would dictate the number of ensemble members to keep track of during the data assimilation procedure. It is also possible to approximate positive Lyapunov

exponents by keeping track of the total expansion or contraction gained along the forecast steps.

4.3 Properties of the EKF for hyperbolic systems with one-dimensional unstable spaces

Let $f : M \circlearrowleft$ be a diffeomorphism having a one-dimensionally unstable hyperbolic attractor $\mathcal{A}$, as in §4.2.1.1. We show that for generic C^2 function h, as in §4.2.1.2, the 2-member ensemble Kalman filter (2MEKF) is $\mathcal{O}(\epsilon)-$reliable, in the sense of Definition 1.

4.3.1 Evolution equations

First, we write down the equations for the 2MEKF, following [HKS07]. See also the simplification discussed in §4.4.1.

Let $\epsilon > 0$. Starting from an initial ensemble of analysis vectors $\overline{x}^a_{n-1} \pm v^a_{n-1}$, we obtain the new background vectors at step n by forecasting according to f. We denote these background ensemble vectors by $\overline{x}^b_n \pm v^b_n$. Thus,

$$\overline{x}^b_n = \frac{1}{2}(f(\overline{x}^a_{n-1} + v^a_{n-1}) + f(\overline{x}^a_{n-1} - v^a_{n-1})),$$
$$v^b_n = f(\overline{x}^a_{n-1} + v^a_{n-1}) - \overline{x}^b_n.$$

The average value of the corresponding observation will be $\overline{h}_n := \frac{1}{2}(h(\overline{x}^b_n + v^b_n) + h(\overline{x}^b_n - v^b_n))$. The measurement of h from the true trajectory x_n will be denoted by $y_n = h(x_n)$. Let $q_n := \frac{1}{\epsilon}(h(\overline{x}^b_n + v^b_n) - \overline{h}_n)$. The corresponding analysis vectors,

obtained using an ensemble square root filter, are $\overline{x}_n^a \pm v_n^a$, with:

$$\overline{x}_n^a = \overline{x}_n^b + \frac{2q_n}{1+2q_n^2}(y_n - \overline{h}_n)\frac{v_n^b}{\epsilon}, \qquad (*)$$
$$v_n^a = \frac{1}{\sqrt{1+2q_n^2}} v_n^b.$$

4.3.2 Basic definitions and notation

Under Takens' genericity conditions (see Theorem 4.2.1.2), it is ensured that for every $x \in M$, the vectors

$$\{\nabla h(x), (Df_x)^T \nabla h(f(x)), (Df_x^2)^T \nabla h(f^2(x)), \ldots, (Df_x^{2N})^T \nabla h(f^{2N}(x))\}$$

span $T_x M$, for all $x \in M$.

Let $\tilde{\gamma}(x) := \frac{1}{\|\nabla h(x)\| |\cos \measuredangle(\nabla h(x), E_x^u)|} = \frac{1}{|(v_x^u)^T \nabla h(x)|}$, where $v_x^u \in E_x^u$ is a unit length vector. We note that $\tilde{\gamma}(x) < \infty$ whenever $\nabla h(x)$ and v_x^u are not orthogonal. By compactness of $\mathcal{A}$, there is some constant $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} > \sup_{x \in \mathcal{A}} \min_{j=0,\ldots,2N} \{\tilde{\gamma}(f^j(x))\}.$$

We note that $\tilde{\gamma}$ is finite by the non-degeneracy condition on h. Whenever $\tilde{\gamma}(x) < \tilde{\gamma}$ we will say that the angle $\measuredangle(\nabla h(x), E_x^u)$ is *good*.

Using the Taylor expansion of h around $\overline{x}_n^b$, we know that for $\|v_n^b\|$ small,

$$h(\overline{x}_n^b \pm v_n^b) = h(\overline{x}_n^b) \pm \nabla h(\overline{x}_n^b) \cdot v_n^b + \mathcal{O}(\|v_n^b\|^2).$$

Hence, $\overline{h}_n = h(\overline{x}_n^b) + \mathcal{O}(\|v_n^b\|^2)$ and $\epsilon q_n = \nabla h(\overline{x}_n^b) \cdot v_n^b + \mathcal{O}(\|v_n^b\|^2)$.

We have assumed the metric is adapted, and $\lambda > 1 > \mu$ are strict lower and upper bounds on the expansion, respectively contraction, along unstable and stable

spaces. Then, whenever x and y are sufficiently close we have, $\|f(x) - f(y)\|_u \geq \lambda\|x - y\|_u$ and $\|f(x) - f(y)\|_s \leq \mu\|x - y\|_s$, where $\|\|_{u(s)}$ denote distance along unstable (stable) spaces, to be defined precisely in the sequel. Let $\overline{\Lambda}$ be a Lipschitz constant for f, and L a Lipschitz constant for h.

4.3.3 Outline of inductive estimates

Here we introduce some further notation aiming to outline the ideas behind the inductive arguments in the coming sections. The new notation in this section is not required in subsequent sections. For any n, let

$$\begin{aligned}
\mathcal{A}_n &:= \frac{\measuredangle(v_n^a, E_{x_n}^u)}{\epsilon}, \\
\mathcal{V}_n &:= \frac{\|v_n^a\|}{\epsilon}, \\
\mathcal{X}_n &:= \frac{\|x_n - \overline{x}_n^a\|}{\epsilon}, \\
\mathcal{S}_n &:= \frac{\|x_n - \overline{x}_n^a\|_s}{\epsilon^2}.
\end{aligned}$$

Relevant properties of the ensemble can be expressed in terms of the above quantities. For example the shadowing property is equivalent to $\{\mathcal{X}_n\}_{n\in\mathbb{N}}$ being bounded. The ensemble size is bounded provided $\{\mathcal{V}_n\}_{n\in\mathbb{N}}$ is bounded.

We will later show that the following inequalities hold, provided $\|x_n - \overline{x}_n^a\|, \|v_n^a\|$ and $\measuredangle(v_n^a, E_{x_n}^u)$ are sufficiently small. First,

$$\mathcal{V}_{n+1} \leq \begin{cases} \frac{\tilde{\gamma}}{\sqrt{2}} & \text{if } \measuredangle(\nabla h(x_{n+1}), E_{x_{n+1}}^u) \text{ is good,} \\ \overline{\Lambda}\mathcal{V}_n & \text{if } \measuredangle(\nabla h(x_{n+1}), E_{x_{n+1}}^u) \text{ is bad.} \end{cases}$$

Thus, $\{\mathcal{V}_n\}$ remains bounded if the number of consecutive bad angles is bounded above. Next, there exist some $0 < \nu < 1$ and $C, C' > 0$, depending on f and h, such

that

$$\mathcal{A}_{n+1} \leq \nu \mathcal{A}_n + C\mathcal{X}_n + C'\mathcal{V}_n.$$

There exist some $0 < \mu < 1$ and $C, C' > 0$, depending on f and h, such that

$$\mathcal{S}_{n+1} \leq \mu \mathcal{S}_n + C\mathcal{A}_n\mathcal{V}_n\mathcal{X}_n + C'\mathcal{V}_n^2.$$

In general, for $\mathcal{X}_n$ we only have

$$\mathcal{X}_{n+1} \leq \overline{\Lambda}(1 + C\mathcal{V}_n)\mathcal{X}_n.$$

These estimates provide some insight on the evolution of the quantities $\mathcal{A}_n$, $\mathcal{V}_n$, $\mathcal{S}_n$, $\mathcal{X}_n$ with respect to n. However, showing that $\mathcal{X}_n$ is bounded requires some further considerations. It is in fact fruitful to study the quantities $\mathcal{Q}_n := \frac{\|x_n - \overline{x}_n^a\|}{\|v_n^a\|}$ instead of $\mathcal{X}_n$. For appropriate choices of the initial ensemble, the size of the perturbation elements in the Kalman filter somehow keeps track of the the distance to the truth. Indeed, when good angles $\measuredangle(E^u_{x_{n+1}}, \nabla h(x_{n+1}))$ occur, $\mathcal{Q}_{n+1} \leq 1 + \frac{\mathcal{Q}_n}{\sqrt{1+2\frac{C_1^2}{\tilde{\gamma}^2}}}$ provided the smallness assumptions above. In fact, contraction by a factor arbitrarily close to $\frac{1}{\sqrt{1+2\frac{C_1^2}{\tilde{\gamma}^2}}}$ occurs provided $\mathcal{Q}_n$ is not too small. When bad angles occur, there may be exponential growth of the quotient $\frac{\mathcal{Q}_{n+1}}{\mathcal{Q}_n}$, but if the number of consecutive bad angles is bounded above, this growth rate can be controlled in such a way that the expansion is compensated by the contraction gained by the occurrence of a good angle. These arguments are enough to show that $\{\mathcal{Q}_n\}$ is bounded, and furthermore, that it is eventually of order one. The same conclusion holds for $\{\mathcal{A}_n\}$ and $\{\mathcal{S}_n\}$.

In the next section, we present an inductive scheme making the above estimates rigorous. It is valid for $\|v_0^a\| = \mathcal{O}(\epsilon)$, and the quantities $\|x_0 - \overline{x}_0^a\|$ and $\measuredangle(v_0^a, E^u(x_0))$

sufficiently small. As we will see in Proposition 4.3.1, in this setting, the shadowing property is guaranteed.

4.3.4 Inductive scheme

We will now establish the fundamental properties of the 2MEKF generated by $f : M \circlearrowleft$ using an inductive scheme. In short, the 2MEKF will be $\mathcal{O}(\epsilon)-$reliable provided the ensemble has been initialized in such a way that the displacement vector v_0 lies in a sufficiently narrow unstable cone and that the distance from the ensemble mean to the true trajectory is sufficiently small.

Proposition 4.3.1 (Properties of 2-member ensemble Kalman filter)**.**

Let $x_0 \in \mathcal{A}$. Then, the following holds.

- For generic observation function h, there is a family $\{\mathcal{I}_\epsilon\}_{\epsilon>0}$ of open sets of initial ensembles such that whenever $\mathcal{E}_\epsilon \in \mathcal{I}_\epsilon$, the 2MEKF with initial ensembles $\{\mathcal{E}_\epsilon\}_{\epsilon>0}$ and noiseless observation of h is $\mathcal{O}(\epsilon)-$reliable. More precisely, this is the case for all 2MEKF initialized in such a way that the inductive hypothesis from §4.3.4.1 holds for suitable choice of constants $C_1, \ldots, C_5$. Moreover, the ensemble spread remains proportional to ϵ.

- The same conclusion holds if the measurement of h has noise, provided its size is bounded by a small multiple of ϵ, and also when the observations are generated by a pseudo-trajectory, provided its distance to a true trajectory is sufficiently small.

The proof of these results occupies the remainder of this subsection.

4.3.4.1 Inductive hypothesis $IH(C_1, C_2, C_3, C_4, C_5, \epsilon)$

Definition 2. Let $\epsilon > 0$. We say that the ensemble with mean $\overline{x}_n^a$ and perturbations v_n^a (or concisely, the ensemble at time n) satisfies the inductive hypothesis $IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon)$ if the following holds:

$(i)_n$ Lower bound on spread of ensemble.

$$C_1\epsilon \leq \|v_n^a\|.$$

$(ii)_n$ Unstable cone.

$$\measuredangle(v_n^a, E_{x_n}^u) \leq C_2\epsilon.$$

$(iii)_n$ Upper bound on spread of ensemble.

$$\|v_n^a\| \leq C_3\epsilon.$$

$(iv)_n$ Shadowing.

$$\|x_n - \overline{x}_n^a\| \leq C_4\epsilon.$$

$(v)_n$ Bound on distance along stable direction.[4]

$$\|x_n - \overline{x}_n^a\|_s \leq C_5\epsilon^2.$$

4.3.4.2 Inductive step

In this section we show that if $\epsilon > 0$ is sufficiently small and $IH_0(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon)$ is valid at the initial time, then

[4] $\|x_n - \overline{x}_n^a\|_s$ is a shorthand for $\sup_{v \in E_{x_n}^{u\perp}, \|v\|=1} |(x_n - \overline{x}_n^a) \cdot v|$

$IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon)$ will remain valid at all times $n \geq 0$, provided ϵ is sufficiently small and some relations between the constants $C_1, \ldots, C_5$ are satisfied. As for each value of ϵ the set of ensembles for which these conditions are valid contains a non-empty open set, this is enough to prove Proposition 4.3.1.

The letters C, C' will denote positive constants independent of ϵ and $C_1, \ldots, C_5$, but may depend on f and h, and are allowed to change from one appearance to the next. The letter ν will denote a constant between 0 and 1 with the same properties as C. The notation $\mathcal{O}_*$ is similar to the asymptotic $\mathcal{O}$ notation, but the constants involved are allowed to depend on $C_1, \ldots, C_5$ as well.

For the rest of this section we suposse the standing assumption

$$IH_0(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon) \text{ and } IH_m(C_1, C_2, C_3, C_4, C_5, \epsilon) \text{ for all } m \leq n \quad \text{(IH)}$$

is valid for some $n \geq 0$, and some (yet to be determined) constants $C_1, \ldots, C_5$. We will show that $IH_{n+1}(C_1, C_2, C_3, C_4, C_5, \epsilon)$ holds. The proof proceeds by induction provided $C_1, \ldots, C_5$ are chosen appropiately.

Proof of $(iii)_{n+1}$. For any $m \geq 0$ we have

$$\|v^a_{m+1}\| \leq \|v^b_{m+1}\| \leq \overline{\Lambda}\|v^a_m\|.$$

Hence, by the standing assumption (IH), for all $0 \leq n < 2N$ we have that $(iii)_{n+1}$ of $IH_{n+1}(C_1, C_2, C_3, C_4, C_5, \epsilon)$ holds.

For $n \geq 2N$, we observe that when the angle $\measuredangle(E^u_{x_m}, \nabla h(x_m))$ is good and is ϵ sufficiently small, if $(iv)_{m-1}$ holds, then we have

$$\|v^a_m\| = \frac{1}{\sqrt{1 + 2q_m^2}}\|v^b_m\| \leq \frac{\tilde{\gamma}\epsilon}{\sqrt{2}}.$$

By the genericity condition on h a good angle will occur within any $2N+1$ consecutive steps, so we can choose m between $n-2N+1$ and $n+1$. Then, choosing

$$\boxed{C_3 \geq \frac{\bar{\Lambda}^{2N}\tilde{\gamma}}{\sqrt{2}}} \tag{4.1}$$

together with the standing assumption (IH) imply that $(iii)_{n+1}$ holds for ϵ sufficiently small.

Proof of $(i)_{n+1}$. Recall that $\lambda > 1$ is a strict lower bound on the expansion of f along E^u and L is a Lipschitz constant for h. Then, $q_{n+1} \leq L\|v^b_{n+1}\|$. Moreover, if $\|v^a_n\| \geq C_1\epsilon$, and ϵ is sufficiently small we have

$$\|v^a_{n+1}\| \geq \frac{\epsilon}{\sqrt{\epsilon^2 + 2L^2\|v^b_{n+1}\|^2}}\|v^b_{n+1}\| \geq \frac{\epsilon}{\sqrt{\frac{1}{\lambda^2 C_1^2} + 2L^2}}.$$

Then, $(i)_{n+1}$ is guaranteed by choosing C_1 such that

$$\boxed{C_1 \leq \frac{1}{\sqrt{2}L}\sqrt{1-\frac{1}{\lambda^2}}}. \tag{4.2}$$

Proof of $(ii)_{n+1}$. Let us assume $C_2\epsilon$ is sufficiently small. Then,

$$\begin{aligned}
\measuredangle(v^a_{n+1}, E^u_{x_{n+1}}) &= \measuredangle(v^b_{n+1}, E^u_{x_{n+1}}) \\
&\leq \measuredangle(v^b_{n+1}, Df_{\bar{x}^a_n}v^a_n) + \measuredangle(Df_{\bar{x}^a_n}v^a_n, Df_{x_n}v^a_n) + \measuredangle(Df_{x_n}v^a_n, E^u_{x_{n+1}}) \\
&\leq CC_4\epsilon + C_2\nu\epsilon + \mathcal{O}_*(\epsilon^2).
\end{aligned}$$

Hence, $(ii)_{n+1}$ holds, for ϵ sufficiently small, as long as

$$\boxed{\frac{C_4}{C_2} < \frac{1-\nu}{C}}. \tag{4.3}$$

Proof of $(v)_{n+1}$. For the background ensemble, we have

$$\|x_{n+1} - \overline{x}^b_{n+1}\|_s \leq \|x_{n+1} - f(\overline{x}^a_n)\|_s + \|f(\overline{x}^a_n) - \overline{x}^b_{n+1}\| \leq C_5\mu\epsilon^2 + CC_3^2\epsilon^2 + \mathcal{O}_*(\epsilon^3).$$

Therefore,

$$\begin{aligned}
\|x_{n+1} - \overline{x}^a_{n+1}\|_s &\leq \|x_{n+1} - \overline{x}^b_{n+1}\|_s + \|\overline{x}^b_{n+1} - \overline{x}^a_{n+1}\|_s \\
&\leq C_5\mu\epsilon^2 + CC_3^2\epsilon^2 + \|(y_{n+1} - \overline{h}_{n+1})\frac{2q_{n+1}}{1+2q_{n+1}^2}\frac{v^b_{n+1}}{\epsilon}\|_s + \mathcal{O}_*(\epsilon^3) \\
&\leq C_5\mu\epsilon^2 + CC_3^2\epsilon^2 + \frac{C'}{\epsilon}\|x_n - \overline{x}^a_n\|\|v^a_n\|C_2\epsilon + \mathcal{O}_*(\epsilon^3) \\
&\leq C_5\epsilon^2(\frac{CC_3^2 + C'C_2C_3C_4}{C_5} + \mu) + \mathcal{O}_*(\epsilon^3).
\end{aligned}$$

Hence, $(v)_{n+1}$ holds, for ϵ sufficiently small, as long as

$$\boxed{\frac{CC_3^2 + C'C_2C_3C_4}{C_5} < 1 - \mu}. \tag{4.4}$$

Proof of $(iv)_{n+1}$. Let $\tau(n)$ be the number of iterates after the last *good* angle, minus one. We will show

Lemma 4.3.2. There exist some constants σ and $\hat{C}_4$ such that

$$\|x_n - \overline{x}^a_n\| \leq \hat{C}_4\sigma^{\tau(n)}\|v^a_n\|.$$

Remark 4.3.3. Lemma 4.3.2 and the already established property $(iii)_{n+1}$ combined with the fact that good angles occur at least once in every $2N$ consecutive iterates guarantee the shadowing property

$$\|x_n - \overline{x}^a_n\| \leq C_3\hat{C}_4\sigma^{2N}\epsilon =: C_4\epsilon.$$

Proof of Lemma 4.3.2. Applying the triangle inequality to Evolution Equations (*) gives

$$\begin{aligned}\|x_{n+1} - \overline{x}^a_{n+1}\| &\leq \overline{\Lambda}\|x_n - \overline{x}^a_n\|(1 + \frac{C\|v^b_n\|}{\epsilon}) \\ &\leq \overline{\Lambda}(1 + CC_3)\|x_n - \overline{x}^a_n\|.\end{aligned}$$

We consider two cases. Let us fix $K < \frac{C_1}{\overline{\Lambda}(1+CC_3)}$.

Case I $\|x_n - \overline{x}^a_n\| \leq K\epsilon$.

Then $\|x_{n+1} - \overline{x}^b_{n+1}\| \leq \overline{\Lambda}K\epsilon$, and therefore $\|x_{n+1} - \overline{x}^a_{n+1}\| \leq \overline{\Lambda}(1+CC_3)K\epsilon < C_1\epsilon \leq \|v^a_{n+1}\|$. The only restriction on $\hat{C}_4$ imposed by this case is $\hat{C}_4 \geq 1$.

Case II $\|x_n - \overline{x}^a_n\| > K\epsilon$.

In this case, $(v)_n$ implies that $\measuredangle(x_n - \overline{x}^a_n, E^u_{x_n}) = \mathcal{O}_*(\epsilon)$. In view of $(ii)_n$, we also have $\measuredangle(x_n - \overline{x}^a_n, v^a_n) = \mathcal{O}_*(\epsilon)$. Let $\beta_n = \|Df_{x_n}v^u(x_n)\|$, where $v^u(x_n) \in E^u_{x_n}$ is a unit length vector. Then,

$$\begin{aligned}\|v^b_{n+1}\| &= \beta_n\|v^a_n\| + \mathcal{O}_*(\epsilon^2), \\ \|v^a_{n+1}\| &= \frac{\beta_n}{\sqrt{1 + \frac{2q^2_{n+1}}{\epsilon^2}}}\|v^a_n\| + \mathcal{O}_*(\epsilon^2), \\ \|x_{n+1} - \overline{x}^b_{n+1}\| &= \beta_n\|x_n - \overline{x}^a_n\| + \mathcal{O}_*(\epsilon^2).\end{aligned}$$

To estimate $\|x_{n+1} - \overline{x}^a_{n+1}\|$, we consider two further subcases.

Case IIa $|\cos\measuredangle(x_{n+1} - \overline{x}^a_{n+1}, \nabla h(x_{n+1}))| \geq \kappa$, where $\kappa > 0$ is a small constant,

depending on f, h, C_1 and C_3, to be specified later. Then,

$$\begin{aligned}\|x_{n+1}-\overline{x}^a_{n+1}\| &= \frac{\beta_n}{1+2q^2_{n+1}}\|x_n-\overline{x}^a_n\| \\ &+ \beta_n\|x_n-\overline{x}^a_n\|(1-\frac{\cos\measuredangle(x_{n+1}-\overline{x}^a_{n+1},\nabla h(x_{n+1}))}{\cos\measuredangle(v^a_{n+1},\nabla h(x_{n+1}))})+\mathcal{O}_*(\epsilon^2)\\ =&\frac{\beta_n}{1+2q^2_{n+1}}\|x_n-\overline{x}^a_n\|+\mathcal{O}_*(\epsilon^2)\le\frac{\hat{C}_4\sigma^{\tau(n)}}{\sqrt{1+2q^2_{n+1}}}\|v^a_{n+1}\|+\mathcal{O}_*(\epsilon^2).\end{aligned}$$

In particular, in this case $\|x_{n+1}-\overline{x}^a_{n+1}\| \le \hat{C}_4\sigma^{\tau(n)}\|v^a_{n+1}\|+\mathcal{O}_*(\epsilon^2)$. Furthermore, $\|x_{n+1}-\overline{x}^a_{n+1}\| \le \frac{\hat{C}_4\sigma^{\tau(n)}}{\sqrt{1+2\frac{C_1^2}{\tilde{\gamma}^2}}}\|v^a_{n+1}\|+\mathcal{O}_*(\epsilon^2)$ when the angle $\measuredangle(E^u_{x_{n+1}},\nabla h(x_{n+1}))$ is good.

Case IIb $|\cos\measuredangle(x_{n+1}-\overline{x}^a_{n+1},\nabla h(x_{n+1}))|<\kappa$. Then,

$$\begin{aligned}\|x_{n+1}-\overline{x}^a_{n+1}\| &\le \beta_n\|x_n-\overline{x}^a_n\|(1+\frac{C\kappa\|v^b_{n+1}\|}{\epsilon}) \le \beta_n(1+CC_3\kappa)\|x_n-\overline{x}^a_n\| \\ &\le \sqrt{1+2C_3^2\|\nabla h\|^2_\infty\kappa^2}(1+CC_3\kappa)\hat{C}_4\sigma^{\tau(n)}\|v^a_{n+1}\|+\mathcal{O}_*(\epsilon^2).\end{aligned}$$

Choosing $\kappa<\frac{1}{\tilde{\gamma}\|\nabla h\|_\infty}$, ensures that Case IIb implies a *bad* angle $\measuredangle(E^u_{x_{n+1}},\nabla h(x_{n+1}))$. Requiring also that

$$\left(\sqrt{1+2C_3^2\|\nabla h\|^2_\infty\kappa^2}(1+CC_3\kappa)\right)^{4N}<1+2\frac{C_1^2}{\tilde{\gamma}^2},$$

ensures that $(iv)_{n+1}$ holds with $\sigma\ge\sqrt{1+2C_3^2\|\nabla h\|^2_\infty\kappa^2}(1+CC_3\kappa)$, provided ϵ is sufficiently small and $\hat{C}_4\ge 1$. □

The last restriction on the constants $C_1,\dots,C_5$ sufficient for the induction to move forward is therefore

$$\boxed{C_4\ge C_3\sigma^{2N}}. \tag{4.5}$$

Hence, the induction can be carried on by choosing, in that order, constants

C_1, C_3, C_4, C_2 and C_5, satisfying the boxed inequalities. The result holds for sufficiently small ϵ.

Finally, we extend the proof to the case of noisy observation. We remark that this case also covers the situation when the measurements do not come from a true trajectory, but from a pseudo-trajectory, provided its distance to a true trajectory is sufficiently small. Let us assume that the noise in the measurement of h from the true trajectory is bounded by $B\epsilon$. The analysis above remains applicable with minor changes, provided B is sufficiently small. A further subdivision of Case IIa is necessary, depending on whether $|\nabla h(x_{n+1})| \geq H$ or $|\nabla h(x_{n+1})| < H$, for some constant H. (Optimizing the choices of κ and H, to in turn maximize the noise size, B, is possible from the inequalities below.) The restrictions on the size of B and constants $C_1, \ldots, C_5$ can be made explicit by adapting the previous computations to this case, obtaining:

$$\begin{aligned}
\frac{CC_3^2 + C'C_2C_3C_4}{C_5} + \frac{BC_2C_3}{\sqrt{2}C_5} &< 1 - \mu, \\
\max\{2BC_3, \frac{B}{\kappa H}\} &\leq C_1, \\
\left(\sqrt{1 + 2C_3^2\|\nabla h\|_\infty^2\kappa^2}(1 + CC_3\kappa) + \frac{C_3B}{C_1}\right)^{4N} &< 1 + 2\frac{C_1^2}{\tilde{\gamma}^2}, \\
\left(\sqrt{1 + 2C_3^2H^2}(1 + CC_3H) + \frac{C_3B}{C_1}\right)^{4N} &< 1 + 2\frac{C_1^2}{\tilde{\gamma}^2}.
\end{aligned}$$

4.3.5 Achieving the inductive hypothesis

The induction presented in §4.3.4.2 motivates the following definition.

Definition 3. Given $\epsilon, C_1, \ldots, C_5 > 0$, we say that an initial ensemble $\mathcal{E}$ is *attracted*

to $IH(C_1, C_2, C_3, C_4, C_5, \epsilon)$ if for some n, the n-th iterate of $\mathcal{E}$ under the 2MEKF satisfies

$IH_n(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon)$. We say that the initial ensembles $\{\mathcal{E}_\epsilon\}_{\epsilon>0}$ are *attracted to* $IH(C_1, C_2, C_3, C_4, C_5, \epsilon)$ if there exists n such that for all ϵ sufficiently small, the n-th iterate of $\mathcal{E}_\epsilon$ under the 2MEKF satisfies $IH_n(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon)$.

Remark 4.3.4. When $\epsilon > 0$ is sufficiently small and $C_1, \ldots, C_5$ satisfy the boxed inequalities, the inducive arguments from §4.3.4.2 imply that if an ensemble $\mathcal{E}$ is attracted to $IH(C_1, C_2, C_3, C_4, C_5, \epsilon)$, then $IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon)$ holds for all sufficiently large n.

Let $C_1, \ldots, C_5$ be constants for which the induction in §4.3.4.2 is valid, with all boxed inequalities in the proof of Proposition 4.3.1 strict. Then, we have the following.

Proposition 4.3.5. Consider initial ensembles $\{\tilde{\mathcal{E}}_\epsilon\}_{\epsilon>0}$ satisfying $IH(\tilde{C}_1, \tilde{C}_2, \tilde{C}_3\overline{\Lambda}^{-2N}, \tilde{C}_4, \tilde{C}_5, \epsilon)$ for constants that also satisfy the boxed inequalities. Then, for generic h, the ensembles $\{\tilde{\mathcal{E}}_\epsilon\}_{\epsilon>0}$ are attracted to $IH(C_1, C_2, C_3, C_4, C_5, \epsilon)$.

Proof. The proof generalizes that of Proposition 4.3.1. We observe that if $\epsilon > 0$ is sufficiently small, condition (iii) of the inductive hypotesis is attracting in the sense that if a *good* angle occurs at step n (this happens at least once within any $2N+1$ consecutive steps for generic h), then condition (iii) of $IH_n(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon)$ is satisfied, as the proof of Proposition 4.3.1 shows. The rest of the inductive argument remains applicable.

Conditions $(i), (ii), (iv)$ and (v) are also attracting in some sense, but not as simply as (iii). The constants $\tilde{C}_1$, $\tilde{C}_4$, $\tilde{C}_2$ and $\tilde{C}_5$, in that order, can be improved until conditions (i), (iv), (ii) and (v) of $IH_n(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon)$ are achieved, and they are maintained thereafter. Here, we explain how to reach (iv) in detail, as is the most involved, and omit the details of the proofs for the other constants, which use similar ideas.

We go back to the cases presented to establish (iv) in the proof of Proposition 4.3.1. We observe that if the ensemble is in Case I, condition (iv) is valid at the next step. In Case IIb, the quotient $\frac{\|x_{n+1}-\overline{x}^a_{n+1}\|}{\|v^a_{n+1}\|}$ deteriorates with respect to the same quotient at time n by a fixed multiplicative factor that can be controlled by the choice of κ, up to higher order terms in ϵ. In Case IIa, this quotient gets reduced by a factor independent of κ, up to higher order terms in ϵ. Again using the non-degeneracy condition on h, and choosing a sufficiently small value for κ, we can ensure exponentially fast decrease of the quotient $\frac{\|x_{n+1}-\overline{x}^a_{n+1}\|}{\|v^a_{n+1}\|}$, until it gets to order 1. (The exponential decrease occurs along times of good angles. In between, this quotient may deteriorate, but this deterioration is controlled by κ). In particular, condition (iv) of $IH_n(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon)$ is achieved.

The upper bound on ϵ for which this argument applies is determined by higher order terms ignored in the above estimates. It depends on f, h and the values of $\tilde{C}_1, \ldots, \tilde{C}_5$. □

Remark 4.3.6. In fact, when the quantities $\|x_0 - \overline{x}^a_0\|$ and $\measuredangle(v^a_n, E^u_{x_n})$ are small, but much larger than ϵ, the 2MEKF algorithm is still useful. Indeed, the induc-

tive procedure from §4.3.4.2 remains applicable when $IH(C_1, C_2, C_3, C_4, C_5, \epsilon)$ is replaced by $IH(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ with $\epsilon \leq \delta \leq c\sqrt{\epsilon}$, for some $c > 0$, where $IH(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ defined as follows.

Definition 4. Let $\epsilon, \delta > 0$. We say that the ensemble satisfies the inductive hypothesis $IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ at time n if the following holds:

$(i)_n^\delta$ Lower bound on spread of ensemble: $C_1\epsilon \leq \|v_n^a\|$.

$(ii)_n^\delta$ Unstable cone: $\measuredangle(v_n^a, E_{x_n}^u) \leq C_2\delta$.

$(iii)_n^\delta$ Upper bound on spread of ensemble: $\|v_n^a\| \leq C_3\delta$.

$(iv)_n^\delta$ Shadowing: $\|x_n - \overline{x}_n^a\| \leq C_4\delta$.

$(v)_n^\delta$ Bound on distance along stable direction: $\|x_n - \overline{x}_n^a\|_s \leq C_5\delta^2$.

In this case, the proof of Proposition 4.3.5 remains applicable and yields the following.

Corollary 4.1. Let $\epsilon > 0$ be sufficiently small. Assume that $IH_0(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon, \delta)$ holds for some initial ensembles $\{\tilde{\mathcal{E}}_\epsilon\}_{\epsilon>0}$, with constants $C_1, \ldots, C_5$ satisfying the boxed inequalities in §4.3.4.2. Then, there exists some $c > 0$ independent of ϵ such that whenever $\epsilon \leq \delta \leq c\sqrt{\epsilon}$, the forward evolution of $\tilde{\mathcal{E}}_\epsilon$ under the 2MEKF satisfies $IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ for all $n \geq 0$. Moreover, $\tilde{\mathcal{E}}_\epsilon$ is attracted to $IH(C_1, C_2, C_3, C_4, C_5, \epsilon)$. In other words, the 2MEKF with initial ensembles $\{\tilde{\mathcal{E}}_\epsilon\}_{\epsilon>0}$ is $\mathcal{O}(\epsilon)-$reliable.

4.3.6 Lyapunov exponent

A consequence of the properties of 2MEKF presented above is that the positive Lyapunov exponent of the true trajectory can be well approximated using the ensemble.

Let x_0 be any initial condition. Let x_j denote its trajectory, and $z_j := \overline{x}^a_j, \tilde{z}_j := z_j + v^a_j$, with $\overline{x}^a_j, v^a_j$ are as in Equations (*). Let v^u be the unstable direction of x_0, $\chi_n = \frac{1}{n}\log\|Df^n_{x_0}v^u\|$ be the n-th step approximation to the positive Lyapunov exponent of x_0, and $\chi = \lim_{n\to\infty}\chi_n$ the corresponding Lyapunov exponent.

Proposition 4.3.7. Let $\tilde{\chi}_n := \frac{1}{n}\sum_{j=0}^{n-1}\log\frac{\|f(z_j)-f(\tilde{z}_j)\|}{\|z_j-\tilde{z}_j\|}$. Assume that the initial ensemble satisfies $IH_0(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon)$, for constants $C_1, \ldots, C_5$ satisfying the boxed inequalities (4.1)-(4.5). Then, for generic h, we have

$$\|\tilde{\chi}_n - \chi_n\| = \mathcal{O}_*(\epsilon)$$

uniformly on n. In particular, $\|\lim_{n\to\infty}\tilde{\chi}_n - \chi\| = \mathcal{O}_*(\epsilon)$.

Proof. This follows from the shadowing property of 2MEKF (iv), invariance of unstable cones (ii), and the boundedness of 2MEKF (iii), showed in Proposition 4.3.1. □

Remark 4.3.8. In fact, any data assimilation algorithm for which the mean of analysis members shadows the true trajectory, and the displacement vectors have uniformly bounded spread and lie close to the unstable direction also gives a good approximation of the positive Lyapunov exponent of the true trajectory.

4.4 Properties of the EKF with higher dimensional unstable spaces

In this section we generalize the properties of the 2MEKF presented in §4.3 to systems with higher dimensional unstable spaces. For the rest of the section, we let $f : M \circlearrowleft$ be a diffeomorphism having a k^u dimensionally unstable hyperbolic attractor $\mathcal{A}$, as in §4.2.1.1 and $h : M \to \mathbb{R}$ be generic in the sense of §4.2.1.2. In this case, we consider an ensemble Kalman filter with $k > k^u$ members, which is denoted kMEKF.

4.4.1 Simplification in analysis step

Here we show a simplification of the analysis step presented in [HKS07] in the case of scalar observation function. This allows to reduce the computational complexity of the algorithm to quadratic order in k, instead of cubic. The simplification is a consequence of the following simple lemmas, whose content can be traced back, at least, to Potter's work in 1964 [May79, Bie77].

Lemma 4.4.1. Let Q be a $k-$dimensional row vector, and let $q^2 = QQ^T$. Then, the symmetric square root of $(I + Q^TQ)^{-1}$ is given by

$$(I + Q^TQ)^{-\frac{1}{2}} = I - \gamma(Q)Q^TQ, \quad \text{where } \gamma(Q) = \frac{1}{q^2}\big(1 \pm \frac{1}{\sqrt{1+q^2}}\big).^{5}$$

Proof. We drop the Q dependence of $\gamma(Q)$ for brevity. Now, we verify the claim directly. First, we note that $I - \gamma Q^TQ$ is symmetric. We let $M = Q^TQ$, and observe

[5] For $Q = 0$, $\gamma(Q) := 0$. To ensure that $I - \gamma(Q)Q^TQ$ is positive definite, the minus sign must be chosen in the definition of $\gamma(Q)$.

that $M^2 = q^2 M$. Thus,

$$(I - \gamma Q^T Q)^2 (I + Q^T Q) = (I - 2\gamma M + \gamma^2 M^2) + (M - 2\gamma M^2 + \gamma^2 M^3)$$
$$= I + (-2\gamma + q^2\gamma^2 + 1 - 2q^2\gamma + q^4\gamma^2) M.$$

The choice of γ ensures that the second term vanishes. □

Lemma 4.4.2. The matrix $W = I - \gamma Q^T Q$ has an orthonormal basis of eigenvectors, with $\frac{1}{\sqrt{1+q^2}}$ as a simple eigenvalue of corresponding eigenvector Q^T and 1 as an eigenvalue with multiplicity $k - 1$.

Proof. The first claim follows from symmetry of W. The second claim can be checked directly. The last claim follows from the fact that for every $v \in \mathbb{R}^k$ such that $v^T Q^T = 0$, $\gamma Q^T Q v = 0$ and thus $Wv = v$. □

4.4.2 Evolution equations

Let $\epsilon > 0$. The evolution equations of the kMEKF are as follows. At time $n-1$ we start with an initial ensemble of vectors with mean $\overline{x}^a_{n-1}$ and displacement vectors $v^a_{j,n-1}$, $1 \leq j \leq k$. To obtain the corresponding background vectors at step n, we forecast according to f. Let us denote the mean of these background ensemble vectors by $\overline{x}^b_n$ and the corresponding displacements by $v^b_{j,n}$. Thus,

$$\overline{x}^b_n = \frac{1}{k} \sum_{j=1}^{k} f(\overline{x}^a_{n-1} + v^a_{j,n-1}),$$

$$v^b_{j,n} = f(\overline{x}^a_{n-1} + v^a_{j,n-1}) - \overline{x}^b_n.$$

The average value of the corresponding observation will be $\overline{h}_n := \frac{1}{k} \sum_{j=1}^{k} h(\overline{x}^b_n + v^b_{j,n})$. The measurement of h from the true trajectory x_n will be denoted by $y_n = h(x_n)$.

For $1 \leq j \leq k$, let $q_{j,n} := \frac{1}{\epsilon}(h(\overline{x}^b_n + v^b_{j,n}) - \overline{h}_n)$. To be consistent with the notation used in [HKS07], we let X^b_n be the matrix whose columns are the displacement vectors $v^b_{j,n}$ and Y^b_n the row vector with entries $\epsilon q_{j,n}$. To make use of the simplification presented in §4.4.1, we let $Q_n = \frac{1}{\sqrt{k-1}\epsilon} Y^b_n$ and $q^2_n = \frac{1}{(k-1)} \sum_{j=1}^k q^2_{j,n} = Q_n Q_n^T$. The corresponding mean and displacement analysis vectors, obtained using an ensemble square root filter, are given by:

$$\begin{aligned}
\overline{x}^a_n &= \overline{x}^b_n + X^b_n\big((k-1)\epsilon^2 I + (Y^b_n)^T Y^b_n\big)^{-1}(Y^b_n)^T(y_n - \overline{h}_n) \\
&= \overline{x}^b_n + X^b_n\big(I - \frac{1}{1+q^2_n}(Q_n)^T Q_n\big)(Y^b_n)^T(y_n - \overline{h}_n) \\
&= \overline{x}^b_n + \frac{1}{(k-1)\epsilon^2}\frac{(y_n - \overline{h}_n)}{1+q^2_n} X^b_n (Y^b_n)^T, \\
X^a_n &= X^b_n (I + Q_n^T Q_n)^{-\frac{1}{2}} = X^b_n (I - \gamma(Q_n) Q_n^T Q_n) =: X^b_n W_n,
\end{aligned}$$

with $\gamma(Q_n) = \frac{1}{q^2_n}\big(1 - \frac{1}{\sqrt{1+q^2_n}}\big)$ for $Q_n \neq 0$ and $\gamma(0) = 0$, as in Lemma 4.4.1. Let

$$v_{0,n} := \frac{1}{(k-1)\epsilon} X^b_n (Y^b_n)^T = \frac{1}{(k-1)} \sum_{j=0}^k q_{j,n} v^b_{j,n}.$$

Then, the equations above simplify to: [6]

$$\begin{aligned}
\overline{x}^a_n &= \overline{x}^b_n + \frac{(y_n - \overline{h}_n)}{1+q^2_n}\frac{v_{0,n}}{\epsilon}, \qquad (**) \\
v^a_{n,j} &= v^b_{n,j} - \gamma(Q_n) q_{j,n} v_{0,n}, \text{ for } 1 \leq j \leq k.
\end{aligned}$$

In words, the coordinates of the displacements of analysis ensemble members from the mean in the ordered basis formed by the background ensemble, i.e. the columns of X^b_n, are given by the columns of W_n, and the transformation from background

[6] When $k = 2$, $q_{1,n} = -q_{2,n}$ and therefore $q^2_n = 2q^2_{1,n}$. Moreover $v_{0,n} = 2q_{1,n} v_{1,n}$. This yields Equations (*).

to analysis ensemble is a contraction by a factor of $\left(1+q_n^2\right)^{-\frac{1}{2}}$ in the direction determined by Q_n (equivalently, by Y_n^b); in model space, this contraction is achieved by a displacement in the direction of $v_{0,n}$.

4.4.3 Inductive scheme

We now generalize the proof of reliability of the 2MEKF to higher dimensions. The main differences between the two cases arise at the linear level. While for systems with one-dimensional unstable spaces the linear analysis is straightforward, the lack of conformality in the forecast step and the fact that at each analysis step there is contraction along (at most) one direction make the inductive step somewhat more challenging in the higher dimensional case.

Adopting a strategy similar to that of §4.3.4 and relying on Takens' embedding theorem proves to be fruitful. In fact, properties (iii) and (v), generalize rather directly. Maintaining a lower bound on the spread of the ensemble in all unstable directions, corresponding to (i), and establishing the shadowing property, corresponding to (iv), require some further work. Property (ii) would remain valid in the setting of $k = k^u + 1$ ensemble members. Here, it is slightly modified to allow for larger ensemble, $k > k^u + 1$.

The main result of the one-dimensional unstable setting, Proposition 4.3.1, is extended to the case of $k^u \geq 1$ unstable directions and $k > k^u$ ensemble members in Proposition 4.4.4.

4.4.3.1 Inductive hypothesis $IH^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$

Definition 5. Given $\epsilon, \delta > 0$, we say that the ensemble satisfies the inductive hypothesis $IH_n^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ at time n if the following holds:

$(i+)$ Lower bound on spread of ensemble:

$$C_1^2 \epsilon^2 \leq \sum_{j=1}^{k} (v \cdot v_{j,n}^a)^2 \quad \forall\, v \in E_{x_n}^{s\perp} \text{ with } \|v\| = 1.$$

$(ii+)$ Closeness of ensemble perturbations to unstable directions:

$$\sum_{j=1}^{k} (v \cdot v_{j,n}^a)^2 \leq C_2^2 \delta^4 \quad \forall\, v \in E_{x_n}^{u\perp} \text{ with } \|v\| = 1.$$

$(iii+)$ Upper bound on spread of ensemble:

$$\|v_{j,n}^a\| \leq C_3 \delta, \quad \forall\, 1 \leq j \leq k.$$

$(iv+)$ Shadowing:

$$\|x_n - \overline{x}_n^a\| \leq C_4 \delta.$$

$(v+)$ Bound on distance along stable directions:

$$|(x_n - \overline{x}_n^a) \cdot v| \leq C_5 \delta^2 \quad \forall\, v \in E_{x_n}^{u\perp} \text{ with } \|v\| = 1.$$

Remark 4.4.3. For $k = 2$, the existence of a constant C_1 satisfying $(i+)$ implies the existence of a (possibly different) constant satisfying (i) of §4.3.4.1. Also, $(i+)$

and $(ii+)$ combined yield (ii) of §4.3.4.1. See Remark 4.4.14 for another implication of $(i+)$.

4.4.3.2 Inductive step

Now, we will extend the main result of §4.3, Proposition 4.3.1, to this setting. Assuming the genericity conditions on f and h stated in the beginning of the section, we have the following.

Proposition 4.4.4 (Properties of k-member ensemble Kalman filter)**.** Let $x_0 \in \mathcal{A}$. Then, there is a family $\{\mathcal{I}_\epsilon\}_{\epsilon>0}$ of open sets of initial ensembles such that whenever $\mathcal{E}_\epsilon \in \mathcal{I}_\epsilon$ for all $\epsilon > 0$, the kMEKF with initial ensembles $\{\mathcal{E}_\epsilon\}_{\epsilon>0}$ and noiseless observations of h is $\mathcal{O}(\epsilon)-$reliable.

The same conclusion holds if the measurement of h has sufficiently small noise of order ϵ, and also when the observations are generated by a pseudo-trajectory, provided its distance to a true trajectory is sufficiently small.

Strategy of the proof. As in §4.3.4.2, the proof of Proposition 4.4 follows from an inductive procedure. We will show that there exist constants $C_1, \dots, C_5, c > 0$ such that whenever $\epsilon > 0$ is sufficiently small, $\epsilon \leq \delta \leq c\sqrt{\epsilon}$ and $IH_0^+(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon, \delta)$ holds for some ensembles $\{\mathcal{E}_\epsilon\}_{\epsilon>0}$, then, the forward evolution of $\mathcal{E}_\epsilon$ under the kMEKF with noiseless measurements of h satisfies $IH_n^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ for all $n \geq 0$. The $\mathcal{O}(\epsilon)$ reliability of kMEKF follows as in the case of 2MEKF. We restrict ourselves to the noiseless case, as the extension to the noisy setting is also similar to that of the 2MEKF. To this end, we divide

the proof of Proposition 4.4.4 in several paragraphs, that give the conditions on the constants $c, C_1, \ldots, C_5$ for the induction to follow.

Remark 4.4.5. Constants C_3 and C_1 are independent of the other ones and may be made explicit from our arguments. The remaining relations among constants $C_1, \ldots, C_5$ in this setting are similar to those obtained in §4.3.4.2. In the coming paragraphs we show how to perform the inductive step without obtaining these relations explicitly.

Standing assumption and notation. For the rest of this section we suppose the standing assumption

$$IH_0^+(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon, \delta) \text{ and } IH_m^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta) \text{ for all } m \leq n \tag{IH+}$$

is valid for some $n \geq 0$, and some (yet to be determined) constants $C_1, \ldots, C_5$. We will show that $IH_{n+1}^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ holds. The proof proceeds by induction provided ϵ is sufficiently small, some relation between ϵ and δ holds, and $C_1, \ldots, C_5$ are chosen appropriately.

Before presenting the proof of the inductive step, we introduce some notation and useful remarks.

Definition 6. For each $x \in \mathcal{A}$, let

$$\Gamma(x) := \max_{\|w\|=1} \min_{n' \in \{0,\ldots,2N\}} \frac{1}{|w^T (Df_x^{-n'})^T \nabla h(f^{-n'}(x))|}, \quad \text{and}$$

$$\Gamma := \sup_{x \in \mathcal{A}} \Gamma(x).$$

Remark 4.4.6. We note that for generic h, it happens that $\Gamma(x) < \infty$ for all $x \in M$, because the set $\{w : \|w\| = 1\}$ is compact and by Takens' theorem, $\{(Df_x^{-n'})^T \nabla h(f^{-n'}(x))\}_{0 \leq n' \leq 2N}$ span $T_x M$. Moreover, $\Gamma < \infty$ because $\Gamma(x)$ is continuous in x and $\mathcal{A}$ is compact.

Definition 7. We say that h makes a *good angle* with E_x^u at time n' if

$$\max_{\|w\|=1,\ w \in E_x^u} \frac{1}{|w^T (Df_x^{-n'})^T \nabla h(f^{-n'}(x))|} \leq \Gamma.$$

Remark 4.4.7. Since $\dim E_x^u = k^u$, the definition of Γ combined with elementary orthogonality considerations implies that for any x, there exists a subset of k^u numbers,

$$\{g_1(x) < \cdots < g_{k^u}(x)\} \subset \{0, \ldots, 2N\}$$

such that for each $1 \leq i \leq k^u$, h makes a good angle with E_x^u at time $g_i(x)$.

Remark 4.4.8. Recall that f is a diffeomorphism and for all $n \in \mathbb{Z}$, $Df_x^n E_x^u = E_{f^n(x)}^u$. Because the norm of Df_x^n is uniformly bounded for $x \in \mathcal{A}$ and $|n| \leq 2N$, there is some $\tilde{\Gamma} > 0$ independent of $x \in \mathcal{A}$ such that whenever h makes a *good angle* with E_x^u at time $0 \leq n' \leq 2N$, we have that

$$\max_{\|w\|=1,\ w \in E^u_{f^{-n'}(x)}} \frac{1}{|w^T \nabla h(f^{-n'}(x))|} \leq \tilde{\Gamma}.$$

Definition 8. We say that the angle $\measuredangle(E_x^u, \nabla h(x))$ is *good* if

$$\max_{\|w\|=1,\ w \in E_x^u} \frac{1}{|w^T \nabla h(x)|} \leq \tilde{\Gamma}.$$

(Note that using the standard definition of angle between a vector $v \in \mathbb{R}^N$ and a linear subspace $E \subset \mathbb{R}^N$ to be $\measuredangle(E, v) := \min_{w \in E \setminus \{0\}} \measuredangle(w, v)$, we have that the angle $\measuredangle(E_x^u, \nabla h(x))$ is good exactly when $\frac{1}{\|\nabla h(x)\| |\cos \measuredangle(E_x^u, \nabla h(x))|} \leq \tilde{\Gamma}$.)

Remark 4.4.9. By Remarks (4.4.7) and (4.4.8), in each sequence of $2N+1$ consecutive iterates $f^{-2N}(x), \dots, f^{-1}(x), x$ there are at least k^u good angles, say at times $0 \leq g_1(x) < \cdots < g_{k^u}(x) \leq 2N$. Moreover, the vectors $\{(Df_x^{-g_i(x)})^T \nabla h(f^{-g_i(x)}(x))\}_{1 \leq i \leq k^u}$ may be chosen to be linearly independent.

Proof of $(iii+)_{n+1}$. Specifically, we prove the following.

Proposition 4.4.10 (Bounded ensemble)**.** Let $C_3 \geq \Gamma k^2$, where k is the number of ensemble members and Γ is given by Definition 6. Then, whenever c and ϵ are sufficiently small, independently of n, and $\epsilon \leq \delta < c\sqrt{\epsilon}$, then $\|v^a_{j,n+1}\| \leq C_3\delta$ for all $1 \leq j \leq k$, i.e. $(iii+)_{n+1}$ holds.

The proof of the Proposition 4.4.10 relies on the following lemma.

Lemma 4.4.11. Let ϵ be sufficiently small, and $m \leq n$. Then, there exists $c > 0$ independent of m and n such that whenever $\epsilon \leq \delta < c\sqrt{\epsilon}$ and $\|v^a_{j,m-2N}\| \leq C_3\delta$ for some $m \geq 0$ and all $1 \leq j \leq k$, we have the following. For all $m - 2N \leq n' \leq m$, [7]

(I) $\|(X^a_{n'})^T \nabla h_{n'}\| < \sqrt{k}\epsilon$,

(II) $\|(Df^{2N} X^b_{m-2N} W_{m-2N} \dots W_{n'})^T \frac{(Df^{n'-m})^T \nabla h_{n'}}{\|(Df^{n'-m})^T \nabla h_{n'}\|}\| < \frac{k\epsilon}{\|(Df^{n'-m})^T \nabla h_{n'}\|}$,

where $\nabla h_{n'}$ is a shorthand for $\nabla h(f^{n'}(x))$. Furthermore,

(III) For any $m - 2N \leq n' \leq m$,

$$\|(X^a_m)^T \frac{(Df^{n'-m})^T \nabla h_{n'}}{\|(Df^{n'-m})^T \nabla h_{n'}\|}\| \leq \frac{k\epsilon}{\|(Df^{n'-m})^T \nabla h_{n'}\|}$$

[7] If $n' < 0$ the content of (I)-(IV) is meaningless.

(IV)

$$\|(X^a_m)^T\| \le \Gamma k^2 \epsilon.$$

Proof of Lemma 4.4.11. Assume $c > 0$ sufficiently small, $\epsilon \le \delta < c\sqrt{\epsilon}$ and $\|v_{j,m-2N}\| \le C_3\delta$ for some $m \le n$ and all $1 \le j \le k$.

- Proof of (I). Let $m - 2N \le n' \le m$. Then,

$$\begin{aligned}\|(X^a_{n'})^T \nabla h_{n'}\| &= \|\big((X^b_{n'})^T - \gamma(Q_{n'})Q^T_{n'}Q_{n'}(X^b_{n'})^T\big)\nabla h_{n'}\| \\ &= \|(I - \gamma(Q_{n'})Q^T_{n'}Q_{n'})Y^T_{n'}\| + \mathcal{O}(\max_{1\le j\le k}\|v^b_{j,n'}\|^2) \\ &= \frac{\epsilon\sqrt{k-1}}{\sqrt{1+q^2_{n'}}}\|Q^T_{n'}\| + \mathcal{O}(\max_{1\le j\le k}\|v^b_{j,n'}\|^2) < \sqrt{k}\epsilon,\end{aligned}$$

  where the last inequality is valid for sufficiently small $c > 0$.

- Proof of (II). Let $m - 2N \le n' \le m$. Recall that

$$X^b_{n'} = Df_{n'-1}X^a_{n'-1} + \mathcal{O}(\max_{1\le j\le k}\|v^b_{j,n'}\|^2), \text{ and } X^a_{n'} = X^b_{n'}W_{n'},$$

  where $Df_{n'}$ is the linearization of f at the point $\overline{x}^a_{n'}$ and $W_{n'}$ was introduced in §4.4.2. Then, in view of the standing assumption (IH+),

$$X^a_{n'} = Df^{n'-m-2N}_{m-2N}X^b_{m-2N}W_{m-2N}\dots W_{n'} + \mathcal{O}(\max_{1\le j\le k}\|v^b_{j,n'}\|^2).$$

  Then, if $c > 0$ is sufficiently small, the following holds for all sufficiently small $\epsilon > 0$.

$$\begin{aligned}&\|(X^a_{n'})^T\nabla h_{n'}\| < \sqrt{k}\epsilon \quad \Rightarrow \\ &\|(Df^{n'-m-2N}X^b_{m-2N}W_{m-2N}\dots W_{n'})^T\nabla h_{n'}\| < k\epsilon \quad \Rightarrow \\ &\|(Df^{2N}X^b_{m-2N}W_{m-2N}\dots W_{n'})^T(Df^{n'-m})^T\nabla h_{n'}\| < k\epsilon \quad \Rightarrow \\ &\|(Df^{2N}X^b_{m-2N}W_{m-2N}\dots W_{n'})^T\frac{(Df^{n'-m})^T\nabla h_{n'}}{\|(Df^{n'-m})^T\nabla h_{n'}\|}\| < \frac{k\epsilon}{\|(Df^{n'-m})^T\nabla h_{n'}\|}.\end{aligned}$$

- Proof of (III). Since $W_{n'}$ is a (non-strict) contraction for every n', we have that

$$\|(X_m^a)^T \frac{(Df^{n'-m})^T \nabla h_{n'}}{\|(Df^{n'-m})^T \nabla h_{n'}\|}\|$$
$$\leq \|(Df^{2N} X_{m-2N}^b W_{m-2N} \dots W_{n'})^T \frac{(Df^{n'-m})^T \nabla h_{n'}}{\|(Df^{n'-m})^T \nabla h_{n'}\|}\| + \mathcal{O}(\max_{1\leq j\leq k} \|v_{j,n'}^b\|^2).$$

Hence, if $c > 0$ is sufficiently small, for all sufficiently small $\epsilon > 0$ and all $m - 2N < n' < m$ we have that

$$\|(X_m^a)^T \frac{(Df^{n'-m})^T \nabla h_{n'}}{\|(Df^{n'-m})^T \nabla h_{n'}\|}\| \leq \frac{k\epsilon}{\|(Df^{n'-m})^T \nabla h_{n'}\|}.$$

- Proof of (IV). First, note that for all $v, w \in \mathbb{R}^N$ such that $w^T v \neq 0$ we have that $\|v\| = \frac{|v^T \frac{w}{\|w\|}|}{|\cos \measuredangle(v,w)|}$. Using (III) and the definition of Γ we have,

$$\|(X_m^a)^T\| \leq k \max_{0\leq j\leq k} \|v_{j,m}^a\|$$
$$\leq k \max_{0\leq j\leq k} \min_{m-2N\leq n'\leq m} \frac{\|(X_m^a)^T \frac{(Df^{n'-m})^T \nabla h_{n'}}{\|(Df^{n'-m})^T \nabla h_{n'}\|}\|}{|\cos \measuredangle(v_{j,m}^a, (Df^{n'-m})^T \nabla h_{n'})|}$$
$$\leq \max_{\|w\|=1} \min_{m-2N\leq n'\leq m} \frac{k^2\epsilon}{\|(Df^{n'-m})^T \nabla h_{n'}\| |\cos \measuredangle(w, (Df^{n'-m})^T \nabla h_{n'})|}$$
$$= \max_{\|w\|=1} \min_{m-2N\leq n'\leq m} \frac{k^2\epsilon}{|w^T (Df^{n'-m})^T \nabla h_{n'}|} \leq \Gamma k^2 \epsilon.$$

□

Proof of Proposition 4.4.10. Let $C_3 \geq \Gamma k^2$, and assume $\epsilon, c > 0$ are sufficiently small. By the standing assumption (IH+), $\|v_{j,0}^a\| \leq C_3 \overline{\Lambda}^{-2N} \delta$ for all $1 \leq j \leq k$, then $\|v_{j,n}^a\| \leq C_3 \delta$ for all $1 \leq j \leq k$ and $0 \leq n \leq 2N$.

Furthermore, it follows from Lemma 4.4.11 and the choice of C_3 that whenever $\|v_{j,m-2N}^a\| \leq C_3\delta$ for some m and all $1 \leq j \leq k$, then $\|v_{j,m}^a\| \leq C_3\epsilon$, for all $1 \leq j \leq k$. Hence, when $n \geq 2N$, using the standing assumption (IH+) we have that $\|v_{j,n+1}^a\| \leq C_3\epsilon \leq C_3\delta$ for all $1 \leq j \leq k$. □

Proof of $(i+)_{n+1}$. Before presenting the main result of this part, we introduce some notation. For each $m \geq 0$ and $\psi \in T^*_{\overline{x}^b_m} M$, the dual space of $T_{\overline{x}^b_m} M$, let

$$\phi^a_m(\psi) = \frac{\sum_{j=1}^k \psi(v^a_{j,m})^2}{\|\psi\|^{*2}},$$

where $\|\psi\|^* := \sup_{\|v\|=1} |\psi(v)|$. Given a basis $\{w_1, \ldots, w_N\}$ of $T_{\overline{x}^b_m} M$ for which $w_j \in E^u$ for $1 \leq j \leq k^u = \dim E^u$ and $w_j \in E^s$ for $k^u \leq j \leq N$, we consider the dual basis $\{w'_1, \ldots, w'_N\}$ of $T^*_{\overline{x}^b_m} M$, defined by $w'_i(w_j) = \delta_{ij}$. The spaces $E^{u'} = \langle w'_1, \ldots, w'_{k^u} \rangle, E^{s'} = \langle w'_k, \ldots, w'_N \rangle$, are independent of the particular choice of the vectors $w_1, \ldots, w_N$. In fact, the one-to-one correspondence between T^*M and TM induced by the Riemannian metric on M defines a one-to-one correspondence between $E^{u'}$ and $E^{s\perp}$.

We let

$$z^a_m = \inf_{\psi \in E^{u'}_{\overline{x}^b_m} \setminus \{0\}} \phi^a_m(\psi) = \min_{\substack{\psi \in E^{u'}_{\overline{x}^b_m} \\ \|\psi\|^* = 1}} \phi^a_m(\psi) = \min_{\substack{v \in E^{s\perp}_{\overline{x}^b_m} \\ \|v\|=1}} \sum_{j=1}^k (v \cdot v^a_{j,m})^2.$$

Let $\phi^b_m(\psi)$ and z^b_m be defined analogously.

The main result of this part is the following.

Proposition 4.4.12 (Spread of ensemble)**.** Let $C_1 \leq \min\{\left(\frac{\lambda^2-1}{\tilde{M}}\right)^3, \frac{3}{\tilde{L}(1+k\overline{\Lambda}^2)C_3^2}\}$, where where $\lambda > 1$ is a strict lower bound on the expansion along unstable directions, $\overline{\Lambda}$ is a Lipschitz constant for f, k is the number of ensemble members and the constants $\tilde{L}$ and $\tilde{M}$ are defined in the course of the proof. Then, whenever c and ϵ are sufficiently small, independently of n, and $\epsilon \leq \delta < c\sqrt{\epsilon}$, then, $z^a_{n+1} \geq C_1^2 \epsilon^2$, i.e. $(i+)_{n+1}$ holds.

Before the proof, we show two auxiliary estimates.

Lemma 4.4.13. For every $m > 0$,

(1) $z_m^b \geq \lambda^2 z_{m-1}^a$.

(2) For all $\psi \in T_x^*M$, $\phi_m^a(\psi) \geq \frac{1}{1+q_m^2}\phi_m^b(\psi)$.

Proof of (1). For any $\psi \in T^*_{\overline{x}_m^b} M$, linear approximation yields

$$\psi(v_{m,j}^b) = Df^*\psi(v_{m-1,j}^a) + \mathcal{O}_*(\delta^2),$$

where $Df^* : T^*M \circlearrowleft$ is defined by $Df^*\psi(v) = \psi(Dfv)$, for $\psi \in T^*_{f(x)}M, v \in T_xM$. Therefore, when $\psi \neq 0$,

$$\phi_m^b(\psi) = \frac{\|Df^*\psi\|^{*2}}{\|\psi\|^{*2}}\phi_{m-1}^a(Df^*\psi) + \mathcal{O}_*(\delta^3).$$

Let $D'f = (Df^*)^{-1}$, the so-called the co-differential of f. The spaces $E^{u'}$ and $E^{s'}$ are invariant under $D'f$, by the corresponding invariance of E^u and E^s under Df. Moreover, for $\psi \in E_x^{u'}$, we have that

$$\begin{aligned}\|D'f\psi\|^* &= \sup_{w \in E^u_{f(x)}\setminus\{0\}} \frac{|D'f\psi(w)|}{\|w\|} = \sup_{v \in E^u_x\setminus\{0\}} \frac{|D'f\psi(Dfv)|}{\|Dfv\|} \\ &= \sup_{v \in E^u_x\setminus\{0\}} \frac{|\psi(v)|}{\|Dfv\|} < \lambda^{-1} \sup_{v \in E^u_x\setminus\{0\}} \frac{|\psi(v)|}{\|v\|} = \lambda^{-1}\|\psi\|^*.\end{aligned}$$

Thus,

$$\phi_m^b(D'f\psi) = \frac{\|\psi\|^{*2}}{\|D'f\psi\|^{*2}}\phi_{m-1}^a(\psi) + \mathcal{O}_*(\delta^3) > \lambda^2\phi_{m-1}^a(\psi) + \mathcal{O}_*(\delta^3).$$

Thus, if δ is sufficiently small, $z_m^b > \lambda^2 z_{m-1}^a$. □

Proof of (2). We work at step m, and to simplify the notation, we drop the explicit dependence on m. Let $\psi \in T_x^*M$, and let $w \in T_xM$ be the vector such that

$\psi(v) = v \cdot w$. Using Cauchy-Schwartz inequality,

$$\begin{aligned}(v_0 \cdot w)^2 &= \frac{1}{(k-1)^2}(\sum_{j=1}^{k} q_j v_j^b \cdot w)^2 \\ &\leq \frac{1}{(k-1)^2}(\sum_{j=1}^{k} q_j^2)(\sum_{j=1}^{k}(v_j^b \cdot w)^2).\end{aligned}$$

Then, $\gamma(v_0 \cdot w)^2 \leq \frac{1}{(k-1)}(1 - \frac{1}{\sqrt{1+q^2}}) \sum_{j=1}^{k}(v_j^b \cdot w)^2$.

Equations (**) yield

$$\begin{aligned}\sum_{j=1}^{k}(v_j^a \cdot w)^2 &= \sum_{j=1}^{k}(v_j^b \cdot w)^2 + \gamma^2 \sum_{j=1}^{k}(q_j v_0 \cdot w)^2 - 2\gamma \sum_{j=1}^{k}(v_j^b \cdot w)(q_j v_0 \cdot w) \\ &= \sum_{j=1}^{k}(v_j^b \cdot w)^2 + \gamma^2 \sum_{j=1}^{k}(q_j v_0 \cdot w)^2 - 2\gamma(k-1)(v_0 \cdot w)^2 \\ &= \sum_{j=1}^{k}(v_j^b \cdot w)^2 + (k-1)\gamma(v_0 \cdot w)^2(\gamma q^2 - 2) \\ &= \sum_{j=1}^{k}(v_j^b \cdot w)^2 - (k-1)\gamma(v_0 \cdot w)^2(1 + \frac{1}{\sqrt{1+q^2}}) \geq \frac{1}{1+q^2}\sum_{j=1}^{k}(v_j^b \cdot w)^2,\end{aligned}$$

where the last inequality follows from the calculation above. Thus,

$$\phi^a(\psi) \geq \frac{1}{1+q^2}\phi^b(\psi).$$

□

Letting $w = \nabla h$, it is straightforward to get the following.

Corollary 4.2.

$$\sum_{j=1}^{k}(v_j^a \cdot \nabla h)^2 = \frac{1}{1+q^2}\sum_{j=1}^{k}(v_j^b \cdot \nabla h)^2. \tag{4.6}$$

Proof of Proposition 4.4.12. Let $\psi \in E^{u'}_{\overline{x}^b_{n+1}} M$, such that $\|\psi\|^* = 1$. We think of ψ as a horizontal vector, so $\psi(v) = \psi v$ for all $v \in T_{\overline{x}^b_{n+1}} M$; thus $\psi\psi^T = 1$. Let

$Dh = Dh^u + Dh^s$ with $Dh^u \in E^{u'}$ and $Dh^s \in E^{s'}$. Let us decompose $\psi = \psi_h + \psi_0$, where $\psi_h \in \langle Dh^u \rangle$, and ψ_0 such that $\psi_0 X^b_{n+1}(X^b_{n+1})^T (Dh^u)^T = 0$. Then,

$$\psi X^b_{n+1}(X^b_{n+1})^T \psi^T = \psi_h X^b_{n+1}(X^b_{n+1})^T \psi_h^T + \psi_0 X^b_{n+1}(X^b_{n+1})^T \psi_0^T.$$

Also, by Lemma 4.4.13(2),

$$\psi X^a_{n+1}(X^a_{n+1})^T \psi^T \geq \frac{1}{1+q^2_{n+1}} \psi_h X^b_{n+1}(X^b_{n+1})^T \psi_h^T + \psi_0 X^b_{n+1}(X^b_{n+1})^T \psi_0^T + \mathcal{O}_*(\delta^3).$$

Fix $K \geq 1$ to be determined later.

Case I $\psi_0 \psi_0^T \leq K \psi_h \psi_h^T$.

Then,

$$\begin{aligned}
\phi^a_{n+1}(\psi) &\geq \frac{\phi^b_{n+1}(\psi)}{1+q^2_{n+1}} = \frac{\phi^b_{n+1}(\psi)}{1+\frac{Dh(Dh)^T}{(k-1)\epsilon^2}\phi^b_{n+1}(\psi_h)} + \mathcal{O}_*(\delta) \\
&\geq \frac{\phi^b_{n+1}(\psi)}{1+\frac{Dh(Dh)^T}{(k-1)\epsilon^2}\frac{\phi^b_{n+1}(\psi)}{\psi_h\psi_h^T}} + \mathcal{O}_*(\delta) \\
&\geq \frac{\phi^b_{n+1}(\psi)}{1+\frac{Dh(Dh)^T}{(k-1)\epsilon^2}2(1+K)\phi^b_{n+1}(\psi)} + \mathcal{O}_*(\delta) \\
&\geq \frac{\phi^b_{n+1}(\psi)}{1+\frac{Dh(Dh)^T}{(k-1)\epsilon^2}4K\phi^b_{n+1}(\psi)} + \mathcal{O}_*(\delta).
\end{aligned}$$

Let $\tilde{L} = \frac{4L^2\lambda^2}{(k-1)}$. Then, Lemma 4.4.13(1) yields $\phi^a_{n+1}(\psi) \geq \frac{\lambda^2 z^a_n}{1+\frac{\tilde{L}}{\epsilon^2}K(z^a_n)}$, for δ sufficiently small.

Case II $\psi_0 \psi_0^T > K \psi_h \psi_h^T$.

By Cauchy-Schwartz inequality,

$$2\psi_0\psi_h^T \leq \frac{1}{\sqrt{K}}\psi_0\psi_0^T + \sqrt{K}\psi_h\psi_h^T \leq \frac{2}{\sqrt{K}}\psi_0\psi_0^T.$$

Hence,

$$1 = \psi\psi^T < (1+\frac{1}{\sqrt{K}})^2 \psi_0\psi_0^T.$$

Then,

$$\begin{aligned}\phi^a_{n+1}(\psi) &\geq \psi_0\psi_0^T\phi^b_{n+1}(\psi_0) + \mathcal{O}_*(\delta^3)\\ &\geq \frac{1}{\left(1+\frac{1}{\sqrt{K}}\right)^2}\phi^b_{n+1}(\psi_0) + \mathcal{O}_*(\delta^3) \geq \frac{\phi^b_{n+1}(\psi_0)}{1+\frac{3}{\sqrt{K}}} + \mathcal{O}_*(\delta^3).\end{aligned}$$

Hence, for δ sufficiently small, Lemma 4.4.13(1) yields $\phi^a_{n+1}(\psi) \geq \frac{\lambda^2 z^a_n}{1+\frac{3}{\sqrt{K}}}$.

Choosing $K = \left(\frac{3\epsilon^2}{\tilde{L}z^a_n}\right)^{\frac{2}{3}}$ and $\tilde{M} = (9\tilde{L})^{\frac{1}{3}}$ yields $\phi^a_{n+1}(\psi) \geq \frac{\lambda^2 z^a_n}{1+\tilde{M}(\frac{z^a_n}{\epsilon^2})^{\frac{1}{3}}}$ whenever $K \geq 1$.

When $K < 1$, $\frac{3\epsilon^2}{\tilde{L}z^a_{n-1}} < 1$ and therefore $z^a_{n+1} \geq \frac{1}{1+q^2_{n+1}} z^a_n \geq \frac{3\epsilon^2}{\tilde{L}(1+k\bar{\Lambda}^2)C_3^2}$.

Thus, $z^a_n \geq \min\{z_0, \left(\frac{\lambda^2-1}{\tilde{M}}\right)^3\epsilon^2, \frac{3\epsilon^2}{\tilde{L}(1+k\bar{\Lambda}^2)C_3^2}\}$. By the standing assumption (IH+) and the choice of C_1, the proof is complete. □

Remark 4.4.14. Condition $(i+)_n$ implies that there is a constant $\tilde{C} > 0$ such that when the angle $\measuredangle(\nabla h_{x_n}, E^u_{x_n})$ is good and δ is sufficiently small, $q_n \geq \tilde{C}$.

Proof. Let $Dh_{x_n} = Dh^u_{x_n} + Dh^s_{x_n}$ with $Dh^u_{x_n} \in E^{u'}_{x_n}$ and $Dh^s_{x_n} \in E^{s'}_{x_n}$. $Dh^u_{x_n} = 0$ if and only if $\nabla h_{x_n} \perp E^u_{x_n}$. By Definition 8 of good angle, there exists some $C > 0$ such that when the angle $\measuredangle(\nabla h_{x_n}, E^u_{x_n})$ is good, $\|Dh^u_{x_n}\|^* > C$. In this case,

$$\begin{aligned}q_n^2 &= \frac{1}{\epsilon^2(k-1)}\sum_{j=1}^{k} Dh^u_{x_n}(v^b_{j,n})^2 + \mathcal{O}_*(\delta)\\ &= \frac{1}{\epsilon^2(k-1)}\|Dh^u_{x_n}\|^{*2}\phi^b_n(Dh^u_{x_n}) + \mathcal{O}_*(\delta)\\ &> \frac{1}{\epsilon^2(k-1)}C^2C_1^2\epsilon^2 + \mathcal{O}_*(\delta).\end{aligned}$$

Letting $\tilde{C} = \frac{C^2C_1^2}{k-1}$ yields the claim, for δ sufficiently small. □

Proof of $(ii+)_{n+1}$. The proof of $(ii+)$ follows similarly to that of Proposition 4.4.12. For each $m \geq 0$, we let

$$\hat{z}_m^a = \sup_{\psi \in E^{s'}_{\overline{x}^b_m} \setminus \{0\}} \phi_m^a(\psi) = \max_{\substack{\psi \in E^{s'}_{\overline{x}^b_m} \\ \|\psi\|^* = 1}} \phi_m^a(\psi) = \max_{\substack{v \in E^{u\perp}_{\overline{x}^b_m} \\ \|v\|=1}} \sum_{j=1}^{k} (v \cdot v_{j,m}^a)^2.$$

We define $\hat{z}_m^b$ analogously.

Then, we have the following.

Proposition 4.4.15 (Closeness to unstable directions)**.** There exists some $C_2 > 0$ independent of n such that whenever c and ϵ are sufficiently small, independently of n, and $\epsilon \leq \delta < c\sqrt{\epsilon}$, then $\hat{z}_{n+1}^a \leq C_2^2 \delta^4$, i.e. $(ii+)_{n+1}$ holds.

As above, the proof relies on two auxiliary estimates.

Lemma 4.4.16. For all $m > 0$,

(1) $\hat{z}_m^b \leq \mu^2 \hat{z}_{m-1}^a + \mathcal{O}(C_3^2 \delta^4)$.

(2) For all $\psi \in T_x^* M$, $\phi_m^a(\psi) \leq \phi_m^b(\psi)$.

Proof. The proof of (1) is analogous to that of Lemma 4.4.13(1), we note that the constant in front of the $\mathcal{O}(C_3^2\delta^4)$ error term depends on f and h only. Part (2) follows directly from the proof of Lemma 4.4.13(2). □

Proof of Proposition 4.4.15. Follows directly from Lemma 4.4.16, by choosing C_2 sufficiently large compared to C_3. □

Proof of $(v+)_{n+1}$. The proof of (v+) is entirely analogous to that of (v).

Proof of $(iv+)_{n+1}$. To shorten notation, for each $m \geq 0$, we let $d_m^{a(b)} = x_m - \overline{x}_m^{a(b)}$ be the displacement of the analysis (background) ensemble mean to the truth.

The goal of this part is to show the following.

Proposition 4.4.17 (Shadowing)**.** There is a constant C_4 depending on f, h, C_1 and C_3 and independent of n such that whenever c and ϵ are sufficiently small, independently of n, and $\epsilon \leq \delta < c\sqrt{\epsilon}$, then $\|d_{n+1}^a\| \leq C_4\delta$, i.e. $(iv+)_{n+1}$ holds.

Before presenting the proof we establish some properties of $d_m^{a(b)}$. We will write $d_m^{a(b)}$ in coordinates with respect to the ensemble displacements $v_{m,j}^{a(b)}$, up to small error terms. Thus, let $e_m^{a(b)}$ be such that $d_m^{a(b)} = X_m^{a(b)} e_m^{a(b)} + \mathcal{O}_*(\delta^2)$. (Note that $e = \mathcal{O}_*(\frac{\delta}{\epsilon})$.)

From §4.4.2, we know that $X_{m+1}^b = Df_{x_m} X_m^a + \mathcal{O}_*(\delta^2)$ and $d_{m+1}^b = Df_{x_m} d_m^a + \mathcal{O}_*(\delta^2)$. Then, we get

$$d_{m+1}^b = Df_{x_m} X_m^a e_m^a + \mathcal{O}_*(\delta^2) = X_{m+1}^b e_m^a + \mathcal{O}_*(\frac{\delta^3}{\epsilon}).$$

For the analysis step, from equations (**) we get

$$d_{m+1}^a = d_{m+1}^b + \frac{(y_{m+1} - \overline{h}_{m+1})}{1 + q_{m+1}^2} \frac{v_{0,m+1}}{\epsilon}.$$

For convenience of notation, we now drop the $m+1$ indices for the reminder of this paragraph, writing indices only when necessary. Recall that $y - \overline{h} = \nabla h \cdot d^b + \mathcal{O}_*(\delta^2) = Dhd^b + \mathcal{O}_*(\delta^2)$ and $v_0 = \frac{1}{(k-1)\epsilon} X^b (Y^b)^T = \frac{1}{(k-1)\epsilon} X^b (X^b)^T Dh^T + \mathcal{O}_*(\frac{\delta^3}{\epsilon})$. Then,

$$\begin{aligned} d^a &= d^b - \frac{1}{1+q^2} \frac{1}{(k-1)\epsilon^2} X^b (X^b)^T Dh^T Dhd^b + \mathcal{O}_*(\frac{\delta^4}{\epsilon^2}) \\ &= X^b (I - \frac{1}{1+q^2} \frac{1}{(k-1)\epsilon^2} (X^b)^T Dh^T DhX^b) e^b + \mathcal{O}_*(\frac{\delta^4}{\epsilon^2}) \\ &= X^a W^{-1} (I - \frac{1}{1+q^2} \frac{1}{(k-1)\epsilon^2} (X^b)^T Dh^T DhX^b) e^b + \mathcal{O}_*(\frac{\delta^4}{\epsilon^2}). \end{aligned}$$

Since $Y = DhX^b + \mathcal{O}_*(\delta^2)$ and $Q = \frac{1}{\sqrt{k-1}\epsilon}Y^b$, we can write

$$d^a = X^a W^{-1}(I - \frac{1}{1+q^2}(Q)^T Q + \mathcal{O}_*(\frac{\delta^3}{\epsilon^2}))e^b + \mathcal{O}_*(\frac{\delta^4}{\epsilon^2}).$$

We can simplify the last expression by recalling that $W = I - \gamma(Q)Q^T Q$. Straightforward algebra shows that

$$d^a = X^a(I - \frac{1}{q^2}(1 - \frac{1}{\sqrt{1+q^2}})Q^T Q)e^b + \mathcal{O}_*(\frac{\delta^4}{\epsilon^2}).$$

Since $q^2 = QQ^T$, up to an error proportional to $\frac{\delta^4}{\epsilon^2} \leq c^4$, e^a_{m+1} is obtained from e^a_m by contracting in the direction of Q^T by a factor of $\frac{1}{\sqrt{1+q^2}}$, and leaving all orthogonally complementary directions unchanged. From Corollary 4.2, we know that in model space, the analysis step contracts the total projection onto ∇h by a factor of $\frac{1}{\sqrt{1+q^2}}$ by means of adjusting in the direction of $X^b(Y^b)^T$ (equivalently, v_0).

The argument above shows that, in the linear approximation, the map $e^a_m \mapsto e^a_{m+1}$ is a (non-strict) contraction. Furthermore, using Remark 4.4.14, we have that each time a good angle $\measuredangle(\nabla h_{x_m}, E^u_{x_m})$ occurs, the contraction is by a factor

$$\nu_m \leq (1 + \tilde{C})^{-\frac{1}{2}} := \nu.$$

By Remark 4.4.9, we know that the composition of the $2N+1$ consecutive contractions $e^a_m \mapsto e^a_{m+2N+1}$, includes at least $k^u = \dim E^u$ contractions by at least ν in linearly independent directions. This implies that the composition is a contraction on E^u, though the contraction factor may be close to 1 if the contraction directions are close to being linearly dependent. But for generic h, we can bound the contraction factor away from 1 by compactness, as in Remark 4.4.6.

Proof of Proposition 4.4.17. Let $\mathcal{C}_{n+1}$ be a space spanned by k^u linearly independent directions of contraction of strength at least $\nu = (1+\tilde{C})^{-\frac{1}{2}}$. The existence of such a space is guaranteed by the previous paragraph. Furthermore, the choice can be made in such a way that there exists a (sufficiently large) K independent of x and n for which, up to higher order terms, all vectors inside a cone

$$\mathcal{K}_{n+1} := \{v = v_{\mathcal{C}_{n+1}} + v_{\mathcal{C}_{n+1}^\perp} | v_{\mathcal{C}_{n+1}} \in \mathcal{C}_{n+1}, v_{\mathcal{C}_{n+1}^\perp} \in \mathcal{C}_{n+1}^\perp, \frac{\|v_{\mathcal{C}_{n+1}^\perp}\|}{\|v_{\mathcal{C}_{n+1}}\|} \leq K\}$$

are contracted by at least $\left(1-\frac{1-\nu^2}{1+K^2}\right)^{\frac{1}{2}}$. In particular, that is the case for all $v \in \mathcal{C}_{n+1}$.

When ϵ, δ are sufficiently small, we may incorporate higher order terms to get that the orthogonal projections of e_{n+1}^a to $\mathcal{C}_{n+1}$ are uniformly bounded, say by C_* independent of n. In Proposition 4.4.10 we proved that the columns of the matrix X_{n+1}^a are bounded by $C_3\delta$. Therefore, the orthogonal projections of d_{n+1}^a to $X_{n+1}^a\mathcal{C}_{n+1}$ are bounded by $\tilde{C}_*\delta$, with $\tilde{C}_*$ independent of n.

By the choice of $\mathcal{C}_{n+1}$, $\measuredangle(X_{n+1}^a\mathcal{C}_{n+1}, E_{n+1}^u) = \mathcal{O}_*(\frac{\delta^4}{\epsilon^2})$, and the multiplicative constant is controlled by the choice of c. Hence, if $c > 0$ is sufficiently small, $\measuredangle(X_{n+1}^a\mathcal{C}_{n+1}^\perp, E_{n+1}^u)$ is bounded away from zero independently of n. Combining this with the already established property $(v+)$ yields an upper bound on $\|d_{n+1}^a\|$ proportional to δ and depending on f, h, C_1, C_3 and smallness of ϵ and c. □

4.4.3.3 Improvement to $\mathcal{O}(\epsilon)$ reliability

Let us assume that $IH_0^+(\tilde{C}_1, \tilde{C}_2, \tilde{C}_3\overline{\Lambda}^{-2N}, \tilde{C}_4, \tilde{C}_5, \epsilon, \delta)$ holds for some sufficiently small c and ϵ with $\epsilon \leq \delta < c\sqrt{\epsilon}$, and suitable constants $\tilde{C}_1, \ldots, \tilde{C}_5$. In §4.4.3.2, we have just proved that $IH_n^+(\tilde{C}_1, \tilde{C}_2, \tilde{C}_3, \tilde{C}_4, \tilde{C}_5, \epsilon, \delta)$ remains valid for all $n \geq 0$.

As in the case of one-dimensionally unstable direction, we can improve this result to the following.

Corollary 4.3. Let $C_1', \dots, C_5'$ be any constants for which the induction of §4.4.3.2 applies. Then, there exist $(C_1, \dots, C_5)$ arbitrarily close to $(C_1', \dots, C_5')$ such that if ϵ is sufficiently small, $IH_n^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \epsilon)$ holds for all sufficiently large n, i.e. $IH_n^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \epsilon)$ is attracting.

Proof. It follows from the induction of §4.4.3.2 and the proof of Proposition 4.4.10 that, for all $n \geq 2N+1$, $\|v_{j,n}^a\| \leq C_3\epsilon$ for all $1 \leq j \leq k$, so that $(iii+)$ of $IH_0^+(C_1, C_2, C_3\overline{\Lambda}^{-2N}, C_4, C_5, \epsilon, \epsilon)$ is attracting.

For $(i+)$, we refer to the proof of Proposition 4.4.12. Since the fixed point $z_* := \left(\frac{\lambda^2-1}{\tilde{M}}\right)^3\epsilon^2$ of $F(z) := \frac{\lambda^2 z}{1+\tilde{M}(\frac{z}{\epsilon^2})^{\frac{1}{3}}}$ is a global attractor, we can also conclude that $(i+)$ is attracting.

The remaining properties may be established in a similar manner; see also the proof of Proposition 4.3.5. □

4.4.4 Lyapunov exponents

The kMEKF also allows to approximate the maximal Lyapunov exponent of $f|_{\mathcal{A}}$, as in §4.3.6. For an initial ensemble with mean $\overline{x}_0^a \in \mathcal{A}$ and perturbations X_0^a satisfying the standing assumption (IH+), with constants $C_1, \dots, C_5$ for which the inductive procedure of §4.4.3.2 holds, we have the following.

Proposition 4.4.18.

$$\chi_{\max} := \lim_{n\to\infty} \frac{1}{n} \log \|Df_{x_0}^n\| = \overline{\lim}_{n\to\infty} \frac{1}{n} \log \|X_n^a (W_0 \dots W_n)^{-1}\| + \mathcal{O}(\epsilon).$$

Moreover, when $k = k^u + 1$ we have

$$\chi_{\max} = \overline{\lim}_{n\to\infty} \frac{1}{n} \log \|\left(\Pi_{j=0}^{n} W_j\right)^{-1}\| + \mathcal{O}(\epsilon).$$

Proof. Let us consider matrices $\tilde{D}f_j$ such that for large j, $\tilde{D}f_j$ is $\mathcal{O}(\epsilon)$ close to Df_{x_j} and

$$X_n^a = \left(\Pi_{j=0}^{n} \tilde{D}f_j\right) X_0^a W_0 \dots W_n.$$

The shadowing condition (iv) of §4.4.3.1 yields

$$\chi_{\max} = \overline{\lim}_{n\to\infty} \frac{1}{n} \log \|\Pi_{j=0}^{n} \tilde{D}f_j\| + \mathcal{O}(\epsilon).$$

Moreover, since

$$\Pi_{j=0}^{n} \tilde{D}f_j X_0^a = X_n^a (W_0 \dots W_n)^{-1},$$

and the span of the columns of X_0^a contains a k^u dimensional space inside an unstable cone around $E_{x_0}^u$,

$$\overline{\lim}_{n\to\infty} \frac{1}{n} \log \|\Pi_{j=0}^{n} \tilde{D}f_j\| = \overline{\lim}_{n\to\infty} \frac{1}{n} \log \|X_n^a (W_0 \dots W_n)^{-1}\|.$$

For the second part, let us assume $k = k^u + 1$. We observe that the upper and lower bounds on the ensemble spread (i) and (iii) of §4.4.3.1 yield

$$C\epsilon \|(W_0 \dots W_n)^{-1}\| \leq \|X_n^a (W_0 \dots W_n)^{-1}\| \leq \sqrt{k} C_3 \epsilon \|(W_0 \dots W_n)^{-1}\|.$$

(The upper bound is straightforward. For the lower bound, we use that for each l, (1) W_l has an orthonormal set of eigenvectors, and all but one of them have eigenvalue 1. The remaining one is Q_l^T, whose corresponding eigenvalue is smaller, (2) $(1, \dots, 1) Q_l^T = 0$, and (3) Zero is a singular value of X_l^a of multiplicity one,

corresponding to the fact that $\sum_{j=1}^{k} v_{j,l}^{a} = 0$; the other k^u singular values of X_l^a are greater than or equal to $C\epsilon$, for some constant C independent of l, due to Proposition 4.4.12.)

Hence,

$$\overline{\lim}_{n\to\infty} \frac{1}{n} \log \|\Pi_{j=0}^{n} \tilde{D} f_j\| = \overline{\lim}_{n\to\infty} \frac{1}{n} \log \|(W_0 \dots W_n)^{-1}\|.$$

Combining, we obtain the result,

$$\chi_{\max} = \overline{\lim}_{n\to\infty} \frac{1}{n} \log \|(W_0 \dots W_n)^{-1}\| + \mathcal{O}(\epsilon).$$

□

Remark 4.4.19. In §4.3.6 we gave a simpler approximation to $\chi_{\max}$ based on the expansion of the ensemble during the forecast steps being essentially a scalar process. The approximation above is based instead on accounting for the contraction of the ensemble during the analysis steps. For a multidimensionally unstable system, the latter approach is computationally simpler, making use of the already computed W_j's.

Bibliography

[AAV07] J.F. Alves, V. Araújo, and C.H. Vásquez. Stochastic stability of non-uniformly hyperbolic diffeomorphisms. *Stoch. Dyn.*, 7(3):299–333, 2007.

[Aba04] M. Abadi. Sharp error terms and necessary conditions for exponential hitting times in mixing processes. *Ann. Probab.*, 32(1A):243–264, 2004.

[ABS96] P. Ashwin, J. Buescu, and I. Stewart. From attractor to chaotic saddle: a tale of transverse instability. *Nonlinearity*, 9(3):703–737, 1996.

[And01] J.L. Anderson. An Ensemble Adjustment Kalman Filter for Data Assimilation. *Mon. Wea. Rev.*, (129):2884–2903, 2001.

[AT05] V. Araújo and A. Tahzibi. Stochastic stability at the boundary of expanding maps. *Nonlinearity*, 18(3):939–958, 2005.

[Bal00] V. Baladi. *Positive transfer operators and decay of correlations*, volume 16 of *Advanced Series in Nonlinear Dynamics*. World Scientific Publishing Co. Inc., River Edge, NJ, 2000.

[Bal07] V. Baladi. On the susceptibility function of piecewise expanding interval maps. *Comm. Math. Phys.*, 275(3):839–859, 2007.

[BEM01] C.H. Bishop, B.J. Etherton, and S.J. Majumdar. Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Mon. Wea. Rev.*, (129):420–436, 2001.

[Bie77] G.J. Bierman. *Factorization Methods for Discrete Sequential Estimation*, volume 128 of *Mathematics in Science and Engineering*. Academic Press, 1977.

[Bla92] M. L. Blank. Chaotic mappings and stochastic Markov chains. In *Mathematical physics, X (Leipzig, 1991)*, pages 341–345. Springer, Berlin, 1992.

[Bow70] R. Bowen. Markov partitions for Axiom A diffeomorphisms. *Amer. J. Math.*, 92:725–747, 1970.

[Bow75] R. Bowen. *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*. Lecture Notes in Mathematics, Vol. 470. Springer-Verlag, Berlin, 1975.

[BP06] L. Barreira and Y. Pesin. Smooth ergodic theory and nonuniformly hyperbolic dynamics. In *Handbook of dynamical systems. Vol. 1B*, pages 57–263. Elsevier B. V., Amsterdam, 2006. With an appendix by Omri Sarig.

[BS08] V. Baladi and D. Smania. Linear response formula for piecewise expanding unimodal maps. *Nonlinearity*, 21(4):677–711, 2008.

[BvLE98] G. Burgers, P.J. van Leeuwen, and G. Evensen. Analysis Scheme in the Ensemble Kalman Filter. *Monthly Weather Review*, 126(6):1719–1724, 1998.

[BY93] V. Baladi and L.-S. Young. On the spectra of randomly perturbed expanding maps. *Comm. Math. Phys.*, 156(2):355–385, 1993.

[BY08] L. Bunimovich and A. Yurchenko. Where to place a hole to achieve a maximal escape rate. Preprint `arXiv:0811.4438 [math.DS]`, 2008.

[Che02] N. Chernov. *Invariant measures for hyperbolic dynamical systems*, volume 1A, pages 321–407. North-Holland, 2002. A. Katok and B. Hasselblatt, eds.

[CM06] N. Chernov and R. Markarian. *Chaotic billiards*, volume 127 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2006.

[Dol04] D. Dolgopyat. Limit theorems for partially hyperbolic systems. *Trans. Amer. Math. Soc.*, 356(4):1637–1689, 2004.

[DY06] M. Demers and L.S. Young. Escape rates and conditionally invariant measures. *Nonlinearity*, 19(2):377–397, 2006.

[Eve94] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res*, 99:10143–10162, 1994.

[Eve09] G. Evensen. *Data assimilation*. Springer-Verlag, Berlin, second edition, 2009. The ensemble Kalman filter.

[FP08] G. Froyland and K. Padberg. Almost-invariant sets and invariant manifolds – connecting probabilistic and geometric descriptions of coherent structures in flows. Preprint, 2008.

[FPET07] G. Froyland, K. Padberg, M.H. England, and A.M. Treguier. Detection of coherent oceanic structures via transfer operators. *Phys. Rev. Lett.*, 98:224503, 2007.

[GPT+07] F. Ginelli, P. Poggi, A. Turchi, H. Chaté, R. Livi, and A. Politi. Characterizing Dynamics with Covariant Lyapunov Vectors. *Phys. Rev. Lett.*, 99(13):130601, Sep 2007.

[GTH09] C. González-Tokman and B.R. Hunt. Scaling laws for bubbling bifurcations. *Nonlinearity*, 22(11):2607, 2009.

[GTHW] C. González-Tokman, B.R. Hunt, and P. Wright. Approximating invariant densities of metastable systems. To appear in Ergodic Theory and Dynamical Systems.

[HKS07] B.R. Hunt, E.J. Kostelich, and I. Szunyogh. Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Phys. D*, 230(1-2):112–126, 2007.

[HM98] P.L. Houtekamer and H.L. Mitchell. Data Assimilation Using an Ensemble Kalman Filter Technique. *Monthly Weather Review*, 126(3):796–811, 1998.

[HO96] B. R. Hunt and E. Ott. Optimal periodic orbits of chaotic systems occur at low period. *Phys. Rev. E*, (54):328–337, 1996.

[Jen06] O. Jenkinson. Ergodic optimization. *Discrete Contin. Dyn. Syst.*, 15(1):197–224, 2006.

[Kal60] R.E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transaction of the ASME Journal of Basic Engineering*, pages 35–45, March 1960.

[Kal02] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, December 2002.

[Kel82] G. Keller. Stochastic stability in some chaotic dynamical systems. *Monatsh. Math.*, 94(4):313–333, 1982.

[Kel89] G. Keller. Markov extensions, zeta functions, and Fredholm theory for piecewise invertible dynamical systems. *Trans. Amer. Math. Soc.*, 314(2):433–497, 1989.

[KH95] A. Katok and B. Hasselblatt. *Introduction to the modern theory of dynamical systems*, volume 54 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1995. With a supplementary chapter by Katok and Leonardo Mendoza.

[KL99] G. Keller and C. Liverani. Stability of the spectrum for transfer operators. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 28(1):141–152, 1999.

[KL09] G. Keller and C. Liverani. Rare events, escape rates and quasistationarity: some exact formulae. *J. Stat. Phys.*, 135(3):519–534, 2009.

[Liv95] C. Liverani. Decay of correlations for piecewise expanding maps. *J. Statist. Phys.*, 78(3-4):1111–1129, 1995.

[LY73] A. Lasota and J.A. Yorke. On the existence of invariant measures for piecewise monotonic transformations. *Trans. Amer. Math. Soc.*, 186:481–488, 1973.

[LY78] T.Y. Li and J.A. Yorke. Ergodic transformations from an interval into itself. *Trans. Amer. Math. Soc.*, 235:183–192, 1978.

[May79] P.S. Maybeck. Stochastics Models, Estimation, and Control: Introduction, 1979.

[MDHS06] E. Meerbach, E. Dittmer, I. Horenko, and C. Schütte. *Multiscale modelling in molecular dynamics: biomolecular conformations as metastable states*, chapter Computer Simulations in Condensed Matter: From Materials to Chemical Biology. Volume I, pages 475–497. Number 703 in Lecture Notes in Physics. Springer, 2006.

[MSF05] E. Meerbach, C. Schütte, and A. Fischer. Eigenvalue bounds on restrictions of reversible nearly uncoupled Markov chains. *Linear Algebra Appl.*, 398:141–160, 2005.

[OHS+04] E. Ott, B.R. Hunt, I. Szunyogh, A.V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D.J. Patil, and J.A. Yorke. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus. Series A: Dynamic Meteorology and Oceanography.*, 56(5):415–428, Oct 2004.

[Ose68] V.I. Oseledec. A multiplicative ergodic theorem. Characteristic Ljapunov, exponents of dynamical systems. *Trudy Moskov. Mat. Obšč.*, 19:179–210, 1968.

[PW97] Y. Pesin and H. Weiss. A multifractal analysis of equilibrium measures for conformal expanding maps and Moran-like geometric constructions. *Journal of Statistical Physics*, 86(1-2):233–275, 1997.

[RGYU99] K. Reif, S. Gunther, E. Yaz, and R. Unbehauen. Stochastic stability of the discrete-time extended Kalman filter . *Automatic Control, IEEE Transactions on*, 44(4):714 –728, apr 1999.

[Rue98] D. Ruelle. General linear response formula in statistical mechanics, and the fluctuation-dissipation theorem far from equilibrium. *Phys. Lett. A*, 245(3-4):220–224, 1998.

[Sin68] Ya. G. Sinai. Markov partitions and C-diffeomorphisms. *Funct. Anal. Appl.*, 2(1):61–82, 1968.

[SN79] I. Shimada and T. Nagashima. A Numerical Approach to Ergodic Problem of Dissipative Dynamical Systems. *Progress of Theoretical Physics*, 61:1605–1616, June 1979.

[SYC91] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *Journal of Statistical Physics*, 65:579–616, November 1991.

[TAB+03] M.K. Tippett, J.L. Anderson, C.H. Bishop, T.M. Hamill, and J.S. Whitaker. Ensemble Square Root Filters. *Monthly Weather Review*, 131(7):1485–1490, 2003.

[Tak81] F. Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980 (Coventry, 1979/1980)*, volume 898 of *Lecture Notes in Math.*, pages 366–381. Springer, Berlin, 1981.

[Vás07] C.H. Vásquez. Statistical stability for diffeomorphisms with dominated splitting. *Ergodic Theory Dynam. Systems*, 27(1):253–283, 2007.

[VHO+96] S. C. Venkataramani, B. R. Hunt, E. Ott, D. J. Gauthier, and J. C. Bienfang. Transitions to bubbling of chaotic systems. *Phys. Rev. Lett.*, 77(27):5361–5364, Dec 1996.

[Via97] M. Viana. Stochastic dynamics of deterministic systems. *Lecture Notes XXI Bras. Math. Colloq. IMPA, Rio de Janeiro*, 1997.

[WBJ04] X. Wang, C.H. Bishop, and S.J. Julier. Which Is Better, an Ensemble of Positive-Negative Pairs or a Centered Spherical Simplex Ensemble? *Monthly Weather Review*, 132(7):1590–1605, 2004.

[WH02] J.S. Whitaker and T.M. Hamill. Ensemble Data Assimilation without Perturbed Observations. *Mon. Wea. Rev.*, 130:1913–1924, 2002.

[YH99] G. Yuan and B. R. Hunt. Optimal orbits of hyperbolic systems. *Nonlinearity*, 12(4):1207–1224, 1999.

[ZHO03] A. V. Zimin, B. R. Hunt, and E. Ott. Bifurcation scenarios for bubbling transition. *Phys. Rev. E*, 67(1):016204, Jan 2003.